An exploration of ensemble visual processing through perception, attention and memory

by
Hee Yeon Im

A dissertation submitted to Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy

Baltimore, Maryland

October 2013

Abstract

This dissertation examines the functioning of ensemble representations at different levels of human visual processing: perception, attention, and memory. An ensemble representation refers to a global presentation that provides a general description of a set of items, computed from multiple individual measurements by collapsing across them. In the work presented in this dissertation, I demonstrate that 1) ensembles are extracted without representing each of individual measurements, similar to visual textures or perceptual grouping, 2) each ensemble representation can function as a unit for selection just as a more familiar construct for selection, i.e., an object, and that 3) each ensemble is encoded and stored in visual working memory in an all-or-none manner.

Committee Members: Dr. Justin Halberda [Advisor]

Dr. Howard Egeth

Dr. Jonathan Flombaum

Dr. Barbara Landau

Dr. Soojin Park

Alternate Committee: Dr. Lisa Feigenson

Dr. Steven Hsiao

Preface

Acknowledgements

This dissertation would not have been possible without the support of many people. Many thanks to my advisor, Justin Halberda, who have provided exceptional support both intellectually and emotionally for five years in Hopkins. He has been instrumental in the development of my thinking and his comments and suggestions have never failed to be insightful and inspiring. Also thanks to my committee members for offering guidance and support. I am very grateful to have talented colleagues around me who are always friendly and fun to be with in the lab and department. Last but certainly not least, thanks to my loving family (both in the US and in Korea) and numerous friends (also both in the US and in Korea) who endured this long process with me, always offering great support and love.

Table of Contents

List of Figures

CHAPTER 1. INTRODUCTION

In this dissertation, I investigate the functioning of "ensemble features" in human visual processing. The terms ensemble and ensemble features are relatively new terms in the vision sciences literature. They are meant to indicate a type of summary representation – like noticing the average size among a group of circles rather than focusing on each of the individual sizes. Because these terms are relatively new, and because there is not yet a vast literature in which previous authors have explored all of the relevant distinctions, I will endeavor through this dissertation to explore the functioning of ensemble representations at many levels throughout visual processing.

The dissertation is structured as a series of chapters that each explores ensemble representations at a particular level of visual processing. For example, In Chapter 2 I will explore ensemble representations in perception and explore a parallel between ensemble processing and texture processing. And, in Chapter 5, I will explore the encoding and storage of ensemble information in visual working memory. A review of the literature relevant to each chapter appears at the beginning of each chapter. Here, I begin with an overview of my dissertation – in Section 1.1 I will briefly introduce the topic of ensemble representations and review the existing literature on ensemble representations that is relevant to all of my chapters; and, in Section 1.2, I will present an overview the experiments included in my dissertation. I have also included a Synopsis at the beginning of each chapter. I will use these sections to briefly remind the reader of the goals of the dissertation, placing each chapter in context, and will highlight the main results for that chapter. The main thread of the dissertation can be experienced by reading just the synopses.

1.1 INTRODUCING ENSEMBLE REPRESENTATIONS

An ensemble representation is a global representation that provides a statistical description of a set of items, computed from multiple individual measurements by collapsing across them. For example, people are remarkably efficient and accurate at computing averages, including the mean size (Ariely, 2001; Chong & Treisman, 2003; Im & Chong, 2009), and average orientation (Dakin & Watt, 1997; Parkes, Lund, Angelucci, Solomon, & Morgan, 2001).

An ensemble representation can be any representation that is computed from multiple individual measurements (for review, see Alvarez, 2011). Individual measurements can be collapsed or combined across space or time, based on a particular feature dimension or two (Eisinger, Im, Pailian, & Halberda, 2013; Emmanouil & Treisman, 2008) and provide a single description of a set of individuals. The value of these kinds of representations has been argued to be that, while rapid and approximate, ensemble representations are rich enough to support an understanding of the visual environment, because the environment often consists of collections of similar objects (e.g., faces in crowd, cars in a parking lot). And, even at a more primitive level, natural images often contain remarkable regularities in terms of physical intensities such as luminance or contrast (Brady & Field, 2000). Due to these redundancies and regularities, the visual environment is highly structured and predictable (Kersten, 1987). Thus, forming ensemble representations by capitalizing on this structure and redundancy is rational and efficient. Converging evidence shows that observers are in fact remarkably efficient and accurate at representing ensemble features, including the mean size (Ariely, 2001; Chong & Treisman, 2003), average brightness (Bauer, 2009), average orientation

(Dakin & Watt, 1997; Parkes et al., 2001; Rubenstein & Sagi, 1990), average location of a collection of objects (i.e., centroid; Alvarez & Oliva, 2008), average direction of motion (Williams & Sekuler, 1984), approximate number (Feigenson, 2008; Halberda, Sires, & Feigenson, 2006), average emotion (Haberman & Whitney, 2007a) and identity (de Fockert & Wolfenstein, 2009) of faces. In these various feature dimensions, multiple local measurements can be combined to give rise to a higher-level description of a collection of similar items.

1.2 OVERVIEW OF THE DISSERTATION

In this dissertation I explore ensemble processing through the lenses of perception, attention, and memory. As an overview, my dissertation will demonstrate that 1) ensembles are extracted without representing each individual measurement, similar to visual textures or perceptual grouping, 2) each ensemble representation can function as a unit for selection, similar to the more familiar construct for selection, i.e., an object, and 3) each ensemble is encoded and stored in visual working memory in an all-or-none manner, similar to a slot-like object-based working memory. Here, I provide a description of each chapter.

1.2.1 Overview of Chapter 2

One of the most active controversies throughout the literature on ensemble representations has focused on the question of whether representing ensemble features requires a mechanism distinct from that of representing individual objects. For example, Myczek and Simons (2008) argued that ensemble representations require no mechanism distinct from simply exploiting focused attention and sampling a few individual items from a visual array. However, an alternative possibility is that humans represent ensemble features relying on a distinct mechanism that is more similar to texture perception in which early feature information is pooled across regions without requiring segmentation or sampling of individual objects (Dakin & Watt, 1997; Im & Halberda, 2012; Malik & Perona, 1990; Parkes et al., 2001). Determining whether ensemble representations rely on mechanisms similar to or distinct from those employed when selecting individual items requires that we measure hallmarks or signatures of visual processing that may distinguish these modes of processing. Two such factors are the

internal noise that affects ensemble representation and the sample size that is used when people build these representations. In Chapter 2, I empirically measure the discrimination thresholds for representing the individual size of a single object and representing the average size of multiple objects. By using a variance summation modeling approach, I estimate both the internal noise and the number of samples that support the representation of ensemble average size. Group fits from the variance summation model determined the estimate of 7.0 samples from each display, which exceeds the widely discussed (but not uncontroversial) estimate of a three- to four-item object-based limit of selective attention (Oksama & Hyönä, 2004; Pylyshyn & Storm, 1988). Additionally, the estimate of internal noise for ensemble size was lower than the internal noise for a single object. Taken together, the results in Chapter 2 suggest that the ensemble representation average size relies on a mechanism that is distinct from segmenting individual items. The work in Chapter 2 has been published in Attention, Perception and Psychophysics (Im & Halberda, 2012).

1.2.2 Overview of Chapter 3

Chapter 3 investigates perceptual grouping as one candidate for a mechanism supporting rapid extraction of ensembles from brief visual scenes. I argue that the ensemble feature approximate number can be extracted following perceptual grouping in which elements are clustered into sub-groups in a fast, pre-attentive manner. Experiment B first implements a new approach to modeling how humans define sub-groups of items within a single larger cluster. I have found that human estimates of groups of items are well described by a k-means clustering algorithm that is widely used for image segmentation in computer vision. I present a model that estimates the number of clusters

within images of dot collections with a single free parameter for center-to-center distance among items (i.e., clustering threshold). The model results show that the best-fit clustering threshold – resulting in cluster estimates that converge with human observers' own verbal estimates – was a distance of 4˚. This estimate was stable from as early as 50 msec of display time, and highly consistent across individuals. Thus, grouping items into clusters relies on a fast pre-attentive process that is well fit by k-means clustering. Experiment C further investigates how such grouping may influence numerical estimation of the total number of items in dot arrays. I have found that subjects tend to underestimate the number of dots when the image contains many clusters (as fit by the clustering algorithm). Using the best-fit clustering thresholds of individual subjects, the model reliably predicts the subject's estimates of the number of individual items across a wide range of images. This provides a computational treatment for what has previously been a simple "rule of thumb" in the literature – that humans tend to underestimate. Together, the results reported in Chapter 3 suggest that the ensemble representation approximate number may rely on a mechanism supporting fast, pre-attentive perception of visual clusters affected by bottom-up perceptual grouping within a visual scene.

1.2.3 Overview of Chapter 4

In Chapter 4, I argue that each ensemble collection functions as a single unit for visual indexing and visual selective attention, akin to being an individual object. I focus on a well-known process that requires attention to individuate items: subtizing. In Experiment D, I recorded response times (RT) when subjects were asked to report the number of ensembles and the number of objects in various arrays. I have found that subitizing ensembles results in an almost identical RT function to that of subitizing

individual objects. That is, just as humans tend to be fast and accurate for determining that there are 1, 2, or 3 objects in an array, they are fast and accurate at determining that there are 1, 2, or 3 ensemble clusters in an array. In Experiment E, I further found that the accuracy of subitizing ensembles highly correlated with the accuracy of extracting approximate number from the ensembles. These results together suggest that, under some conditions, ensembles receive a single visual index and function the same way an object does for visual attention – suggesting that ensemble representations, built from multiple samples, serve as individual units for attentional selection.

1.2.4 Overview of Chapter 5

In Chapter 5, I ask whether the precision of ensemble representations improve with increased processing time (e.g., display time) and decreased item load (e.g., number of groups). I present an experiment in which subjects were presented with a flash containing 1, 3, or 6 sets of multiple dots for varying durations and were asked to estimate the number of dots in one of the sets from memory, where the probed set was highlighted after the stimulus flash and a mask disappeared. Because behavioral responses reflect a mixture of responses based on internal representations and those based on strategic guesses, it has become crucial to sort a subject's responses into these two classes (i.e., responses based on an internal representation of the stimulus, and those based on strategic guessing) in order to determine if internal representations are indeed 'flexible' (i.e., with variable precision across time and load) or 'fixed' (i.e., at a specific level of precision). For doing so, I present a novel mixture modeling method that includes both the empirically measured guessing pattern from each individual subject and an empirically-appropriate model for human internal representation of approximate number.

This approach is able to accurately sort human observers' responses into those based on an internal representation of the target and those drawn from their non-random, strategic guessing pattern. Previously, this approach has not been used in the literature and other authors have merely assumed that human guessing is uniform and random. I demonstrate that this assumption is inappropriate.

The empirical guessing pattern was measured for each individual subject by including trials of 0-msec-duration in which no stimulus was actually presented. These 0-msec-duration trials were not noticed by subjects because they were randomly intermixed with other various durations (33 - 198 msec) and every trial was presented with an effective mask. I found that subjects guessing patterns on the 0-msec-duration trials were surprisingly consistent across individuals, and clearly not random nor uniform. It seems that subjects had a strong tendency to choose values in the middle of the response scale (which ranged from 1 to 50) and avoided the values in the extremes, with the overall guessing pattern being unimodal.

Internal representation, the other component of the mixture model, was characterized by a power function between the target feature value and a subject's behavioral response. The key factors that determine the shape of the internal representation in this mixture model are an exponent of the mapping function (β) and the shape of standard deviation (SD) that linearly increases with the target feature value.

The empirically measured guessing pattern and the appropriate functional description for internal representation enables the mixture model to more accurately remove guessing trials from behavioral responses and to provide a more reliable estimate of the precision of internal representations. I found that the mixture model estimates of

the precision (inverse of SD) of number representations at different processing times (33 - 198 msec) and with different item load (1,3, and 6 sets) remained constant, suggesting that the precision of number representation is not flexible but fixed. What changed over processing time and with varying item load instead was the proportion of guessing trials. That is, when processing time is insufficient or when item load is high, subjects do not rely on their internal representations but choose to strategically guess. Together, the results in Chapter 5 suggest that the representation of the ensemble feature approximate number has a fixed precision and that encoding of approximate number into visual working memory is a discrete and all-or-none process.

1.2.5 Summary Statement and Caveat

This dissertation represents an attempt to understand ensemble processes in vision from early perceptual processes (e.g., Chapter 2, textures and sampling; Chapter 3, perceptual grouping), through working memory (e.g., Chapter 5, encoding precision and memory capacity). As I've already highlighted, the term "ensemble representation" is a relatively new construct in the literature. As such, the work in this dissertation must be treated as somewhat of a "case study", in that I focus here primarily on approximate number with some discussion of average size as well. Because the construct of "ensemble representation" might itself turn out to be an umbrella term encompassing a diversity of processes and representations that may not all function alike (e.g., average orientation and speed pooling mechanisms might be implemented in earlier visual areas than approximate number), the work in this dissertation stands as a thorough investigation of approximate number and must be taken with some caution as a case study for how other ensemble representations may function in perception and cognition.

CHAPTER 2. ENSEMBLES IN PERCEPTION, 1: THE ENSEMBLE FEATURE AVERAGE SIZE IS EXTRACTED WITHOUT REPRESENTING INDIVIDUAL SIZES, SIMILAR TO VISUAL TEXTURES

2.1 SYNOPSIS

In Chapter 2, I will argue that ensemble visual processing does not require segmentation of individual objects but rather relies on a global pooling process. The results of Experiment A support this claim by showing that 1) the estimated internal noise affecting ensemble representations is lower than the internal noise of representing individual objects and 2) the estimated number of samples required for ensemble representations is much more than the number of individual objects that can be selected at once.

I begin in Section 2.2 with a review of the relevant literature on texture perception, because the insights from the study of texture processing is informative in understanding how global pooling mechanism can provide ensemble representations without segmenting individual elements. I also review the framework of two different modes of attention (focused vs. distributed) in order to draw a direct analogy to two distinct mechanisms: one for ensemble representation (possibly relying on distributed attention), and one for the representation of individual objects (possibly relying on focused attention).

2.2 BACKGROUND

2.2.1 Texture perception

The nature of texture processing has been extensively explored and reveals the extremely high precision and efficiency of the visual system when extracting texture information from a visual scene within a brief exposure (Bergen & Julesz, 1983; Portilla & Simoncelli, 2000). The existing literature on texture perception already shows that statistical properties of a texture region are effortlessly processed by the visual system (Baker & Mareschal, 2001; Dakin, 2001; Morgan & Glennerster, 1991; Parkes et al., 2001; Victor, Chubb, & Conte, 2005) and affect segmentation of subparts of the visual scene (Bergen & Adelson, 1988; Grossberg & Mingolla, 1985; Malik & Perona, 1990; Wolfson & Landy, 1998). Thus, texture appears to be processed before the segmentation of individual objects.

Global orientation of variable textons has been one of the most prevalent examples of extraction of a global texture attribute (Blake & Marinos, 1990). Global orientation of a textural area seems to result from the integration of local orientation measures. For example, Sagi (1990) found that performance on visual search for a vertical target was affected by the number of horizontal distractors in the search array, suggesting a compulsory integration of horizontal distractors with a vertical target over the entire search array. Such data has supported the notion of "hyper-filters", in which local orientation measures over an entire area are integrated, without segmentation (Sagi, 1990). In addition, segregation of areas by their textures has been shown to be sensitive to the rate of change of texel orientation across space but not to the local density of texels

(Landy & Bergen, 1991). These studies suggest that performance on visual search and texture segregation tasks, based on global orientation, can be explained by mechanisms in which differences between spatially integrated local orientation measures are extracted prior to segmenting individual objects. Such a mechanism involves a further stage of combination and processing of local orientation information in the system, beyond the primitive representation in V1 (Hubel & Wiesel, 1962), but does not necessarily involve the processing of individual objects with bound features that are later averaged.

Processing of the average orientation in a texture has also benefited from study through computational modeling. One promising method of testing models of texture processing is by investigating the effects of noise on texture perception. Dakin and Watt (1997) observed how subjects' performance in perceiving average orientation of textures deteriorates with the addition of external and internal noise. An example of external noise is the variability in the values for a given feature dimension in a scene. For example, the global orientation of the texture for the tree bark in Figure 2.1 is made up of smaller local orientations, many of which are in line with the average orientation (i.e., zero noise) and some of which are a little to the clockwise or counter-clockwise direction of this average; the variance in local orientations throughout the image is the level of external noise.

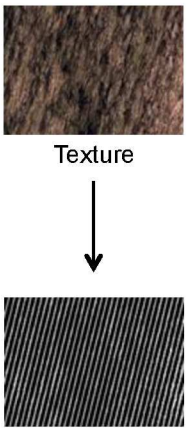


Figure 2. 1 An example of texture from a natural image.

Conversely, internal noise can be understood to derive from the limitations, or inherent noise, of the system when representing a single oriented contour. Additional internal noise could derive from any biases or information loss in the averaging process that combines the local orientation samples. That is, the variance of local orientations is a source of external noise that can limit the perception of the average orientation of a texture in combination with the constant internal noise of the system. Data suggest that integration of texton orientation occurs prior to the judgment of average orientation of a texture field and the arithmetic mean of texton orientations predicts thresholds well (Watt, 1991). In addition to the arithmetic mean, observers can extract the variability of orientations in a texture field (Dakin & Watt, 1997). The integration of orientation seems to require a low attentional load; it can be extracted within a brief exposure of 100 msec to the stimulus. Computational models of texture perception suggest that only a subset of textons is sampled and pooled over the visual array. For example, the variance summation model (Dakin, 2001) fitted to the psychophysical data from human subjects revealed that providing the model with a subset of the textons in the visual array, determined by a power function, could yield the level of discrimination thresholds comparable to those from human subjects' data. The number of samples required by the model increased from 4 to 90 textons while the actual number of textons presented to the humans increased from 4 to 1024, suggesting fairly coarse processing by the human subjects. Indeed, fine-grained representation of individual elements of texture seems to be unnecessary in texture perception as any high-resolution representation of individual textons becomes inaccessible due to the presence of other neighboring textons, just as in the crowding effect in peripheral vision. For example, Parkes et al. (2001) showed that

even when information about the individual orientations of items in the periphery is lost due to crowding by adjacent distractors, the orientation of these items can affect the representation of the average orientation of the array, suggesting that texture processing might occur even without a rigid and stable representation at the level individual objects.

It seems that there may be a strong analogy between ensemble representation and texture perception. It has been suggested that ensemble representation and texture processing may rely on the similar processes (Cavanagh, 2001; Chong & Treisman, 2003; Haberman & Whitney, 2010; Parkes et al., 2001). For example, recent findings on ensemble representation suggested that just as in texture perception (e.g., Parkes et al., 2001), individual representations of elements may be lost or at least be discounted while a more holistic, overall impression is maintained (Alvarez & Oliva, 2009; Haberman & Whitney, 2010). In addition, ensemble representations also appear to be basic visual attributes that are susceptible to perceptual adaptation (e.g., approximate number: Burr & Ross, 2008; mean size: Corbett, Wurnitsch, Schwartz, & Whitney, 2012).

Despite the similarities, a strong connection between ensemble representations and texture perception remains to be made. At the moment, these two lines of research have been conducted separately and differently. For example, the literature on texture perception has focused on an observer's ability to extract average feature information over region of the scene or even over a whole scene, whereas the newer literature on ensemble representations has focused on an observer's ability to extract the statistical average from an array of well-defined, separately attendable objects. In a typical texture stimulus, it is difficult (and sometimes impossible) to recognize the individual elements that make up the texture (Dakin & Watt, 1997; Dakin, 1999; Morgan, Chubb, &

Solomon, 2008) because each element has unclear, blurred edges and many elements are spatially overlapping one another in a display. This is on purpose, because researchers interested in texture have wanted to avoid the possibility of object-based processing in order to focus on texture-based processing. In contrast, the visual stimuli that have been used in studies of ensemble representation typically contain many fewer items (e.g., 4-35 objects) and each item can be easily recognized as a separable object (e.g., dots, faces, and so on) without any spatial overlaps. Thus, while there may be similarities between texture processing and ensemble processing, these connections have yet to be fully explored.

Another complication of the previous research that makes it difficult to draw a direct connection between textures and ensembles is that most of the feature dimensions that the literature on ensemble representation has investigated (e.g., size of circles: Ariely, 2001; or gender, emotion, or identity of faces: Haberman & Whitney, 2007b) do not lend themselves to being understood as basic visual features like orientation of a texton. For example, average size was one of the first ensemble features to be investigated (Ariely, 2001) but it has perhaps drawn the most skepticism and widest criticism - owing, perhaps, to the suggestions that early visual areas have no "size-tuned" neurons (Myczek & Simons, 2008) and that representing object size is traditionally thought to require selecting individual objects from the background (Bundesen & Larsen, 1975; Cave & Kosslyn, 1989). Thus, in a direct denial of a connection to non-object-based texture processing, criticisms of the literature on average size have primarily focused on the possibility that average size is computed from object-based sampling strategies, with focused attention involved. In Chapter 2, I directly address this issue and

present empirical data suggesting that representing average size does not rely on a mechanism that requires segmentation or sampling of individual objects.

Recently it has been suggested that texture perception might also provide an appropriate description for much of everyday visual processing, such as representing the gist of natural scenes or arrays of segmentable objects, and may also operate at higher-level stages of visual processing than previously believed (Freeman & Simoncelli, 2011). For example, visual crowding effects in arrays of individual objects (e.g., in peripheral vision) has been hypothesized to arise from compulsory pooling of peripheral information, just as in texture processing (Parkes et al., 2001; Denis G Pelli, Palomares, & Majaj, 2004). Further, Balas et al. (2009) showed that a texture model based on a statistical description of the visual array (Portilla & Simoncelli, 2000) provided an accurate fit to human performance for the visual crowding effect in peripheral vision in displays containing multiple objects. Together, the insights from the study of texture processing may still be particularly valuable and informative when critically evaluating the newer literature on ensemble representation.

2.2.2 Focused attention vs. Distributed attention

The issue of whether ensemble representation requires segmenting and sampling of individual objects can be addressed under the framework of two different modes of attention: focused attention and distributed attention (for review, see Treisman, 2006). These two modes of attention provide different types of information about the visual scene. As we navigate through the visual environment, for example, we sometimes focus our attention on a single object, such as a tree, and sometimes we spread our attention over a larger area to see only the forest as a whole. We can then represent two different

types of information from the forest: information about the individual trees (e.g., height, color, or orientation of the individual trees) and information about the global properties of the forest on its own (e.g., area, darkness, or density of the whole forest). As in this example, most visual environments can be hierarchically structured. Thus, perceptual mechanisms have evolved to form representations that connect objects hierarchically, and in this way we come to represent properties of the wholes (e.g., forest) as well as properties of their component parts (e.g., trees). Different aspects of visual processing seem to be favored by each mode of attention. At one extreme, we bind the features of a single object and perhaps select and identify it using focused attention on the single object (Kahneman, Treisman, & Gibbs, 1992; A. Treisman & Gelade, 1980). At the other extreme, we may obtain rapid access to the gist, a global layout and semantic interpretation of the scene as a whole using globally distributed attention over the scene (e.g., Li, VanRullen, Koch, & Perona, 2002; Potter, 1976), possibly without having to focus on any of the single elements in the scene.

For focused attention, much research has demonstrated that there are severe limits for our capacity to process, segment, and store many individual items in parallel. For example, observers cannot attend or store more than only a few items (e.g., three- or four-item limits in attention and in visual working memory: Alvarez & Franconeri, 2007; Bays & Husain, 2008; Cowan, 2001; Dobkins & Bosworth, 2001; Franconeri, Alvarez, & Enns, 2007; Luck & Vogel, 1997; Palmer, Ames, & Lindsey, 1993; Palmer, 1990; Sakai, Morishita, & Matsumoto, 2007; Simons & Levin, 1997). Some authors have suggested that these limits are imposed because of some fixed amount of overall energy, or resource (e.g., neuronal real estate), available to the brain and by the energy cost of the neuronal

activity that is involved in computation (for review, see Carrasco, 2011). Given such limited amounts of overall energy, so the story goes, stimuli compete for limited resources for further processing (Broadbent, 1958; Neisser, 1967; Treisman, 1960), and this notion is supported by electrophysiological, neuroimaging, and behavioral findings (for review, see Beck & Kastner, 2009; Desimone & Duncan, 1995). When observers attend to a given location in a visual array, competition is biased in favor of the neurons encoding information at the attended area. As a result, neurons with receptive fields at that location become more active, while other activation is suppressed (Desimone & Duncan, 1995). In this way, focused attention allows us to overcome our brain's limited capacity by selectively processing only some part of the available objects or information. Focused attention optimizes the use of the system's limited resources by enhancing the representations of the relevant, while diminishing the representations of the less relevant.

In contrast to focused attention on individual items, Treisman (2006) suggested that ensemble representations may be extracted in the mode of distributed attention. Distributed attention is deployed globally over several objects at once or over the scene as a whole and may not be affected by such severe capacity limitations as focused attention (e.g., as indexed by number of relevant items). Distributed attention was investigated by the classic study pioneered by Navon (1977) in which global and local processing of shape information (e.g., letter shape) competed with each other. In his study, a large letter shape was made of multiples of a different letter shape, each with a smaller size. When observers saw this stimulus, they were better and faster at recognizing the global letter than local letter. Navon reported this finding as the evidence supporting global precedence. Global precedence, in a different sense, is also reflected in the

asymmetry of interference between global and local forms: A global form typically creates greater interference in the processing of a local form than vice versa (Hoffman, 1980; Miller, 1981; Pomerantz, 1983).

2.3 EXPERIMENT A: ESTIMATING OBSERVERS' INTERNAL NOISE AND NUMBER OF SAMPLES FOR EXTRACTING AVERAGE SIZE

Here I present an empirical study designed to determine whether ensemble representation requires segmenting and sampling individual objects. I use a variance summation modeling approach in order to estimate internal noise and sample size for the ensemble process that pools evidence across multiple samples for average size. In this chapter I address two specific questions regarding the mechanisms supporting ensemble representation: is the internal noise for ensemble processing lower than the internal noise for processing individual items, and does the number of samples required by ensemble processing exceed the 3-4 item limit of object-based attention? If so, each of these points would suggest that ensemble representation does not require the segmentation of individual objects.

2.3.1 Experimental method

I relied on a standard two-alternative forced choice psychophysical discrimination task in which subjects had to judge which of two briefly flashed arrays had either the larger individual size or the larger average size. Each of the participants completed two separate sets of experimental blocks for measuring discrimination thresholds of

representing ensemble orientations of multiple gratings and representing individual orientation of a single grating. Having both sets of experimental blocks (Ensembles and Single gratings) allowed for direct comparisons between representations of ensembles and individual objects within subject.

Subjects

16 subjects participated in the experiment. Two among the 16 subjects were experienced subjects and the others were naive subjects. All of the subjects had normal or corrected-to-normal vision. The naive subjects received course extra credit for participating.

Apparatus and stimuli

The stimuli were generated using MATLAB software, together with the Psychophysics Toolbox extensions (Brainard, 1997; D G Pelli, 1997), and were displayed on an LCD monitor driven by a Macintosh iMac computer (the viewable area was a gray central square window with a 17-in. diagonal). The subjects were seated approximately 50 cm from the screen and viewed the display binocularly. At this viewing distance, each pixel was approximately 0.04 of visual angle, and each grating subtended between 1.6 and 4.0 of visual angle. The stimuli were presented on a gray background and consisted of one or more sinusoidal gratings with a spatial frequency of 4 cycles/deg and a Michelson contrast of 99.8%.

In the Ensemble block, multiple gratings (9, 11, 13, 16, 19, or 23 gratings) were randomly located on the display, subtending 56 x 40 of visual angle. In the Single-grating block, only one grating appeared within this same viewing area. The locations of gratings

varied across the two stimulus arrays in order to minimize the masking of stimuli in the second flash by those in the first flash.

To focus on the ensemble processing of distinct objects, we ensured that each item in the array was a segmentable individual object. Each object on the display had a clearly drawn border that was salient in order to avoid blur and reduce the blending of gratings and background. Because crowding by adjacent stimuli occurs in a compulsory manner and may be equivalent to perceiving texture (Parkes et al., 2001), we also ensured that adjacent gratings were separated from each other by at least half of their eccentricity in the display- that is, crowding in foveal vision only occurs over very small distances (2-6 arcmin; Toet & Levi, 1992) whereas crowding in peripheral visual occurs over larger distances, roughly at about half of the eccentricity (Pelli & Tillman, 2008). This ensured that our displays were not crowded.

Procedure

Figure 2.2 illustrates the procedure. The procedure required subjects to view two brief displays (100 ms each), one after the other, and then to judge which display, the first of second, contained either the larger average size (Ensemble blocks) or the larger individual grating (Single-grating blocks). Auditory feedback for errors was provided throughout. The short exposure duration of 100 ms was chosen to prevent scanning eye movements (Morgan, Ward, & Castet, 1998). The interstimulus interval was randomly varied from 1000 to 2200 ms, making the onset of the second flash unpredictable, in order to disrupt strategic planning in preparation for the second display. This delay also reduced any afterimage effects of the first stimulus display, as well as possible effects of the apparent rotation of the second display from the first.

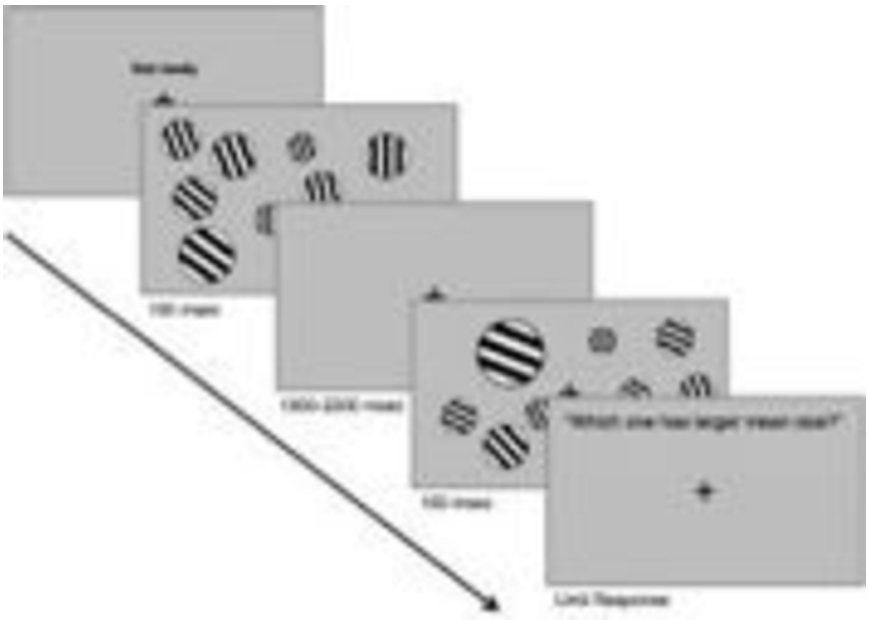


Figure 2. 2 A sample trial of Ensemble block.

We relied on a standard adaptive procedure (Watson & Pelli, 1983) identify each subject's threshold. Six interleaved QUEST routines of 40 trials each were run for each threshold, and the final threshold estimate was obtained by fitting the data with a Weibull function and evaluating thresholds at 75% correct responses. The Weibull function was fitted to the behavioral data using the Psignifit toolbox, Version 3.0, for Python, which implemented the maximum likelihood method described by Wichmann and Hill (2001).

On each trial, the QUEST procedure determined the sizes for the gratings to be displayed in order to determine the size difference that resulted in 75%-correct performance for the subject. There was a standard size and a comparison size for each trial. One of the two displays (first or second) always showed the standard size of 2.8° in diameter.

In the Ensemble block, we systematically varied the external variability of sizes among the items within each ensemble (see Figure 2.3). The size for each grating within

a display was randomly drawn from a Gaussian distribution centered on the size that QUEST had specified for that display (e.g., 2.8° in diameter, or a comparison size) with a standard deviation (σ = 1°, 3°, 6°, or 10°). These SDs were varied across four separate sub-blocks of trials and the order of the sub-blocks was randomized across subjects.

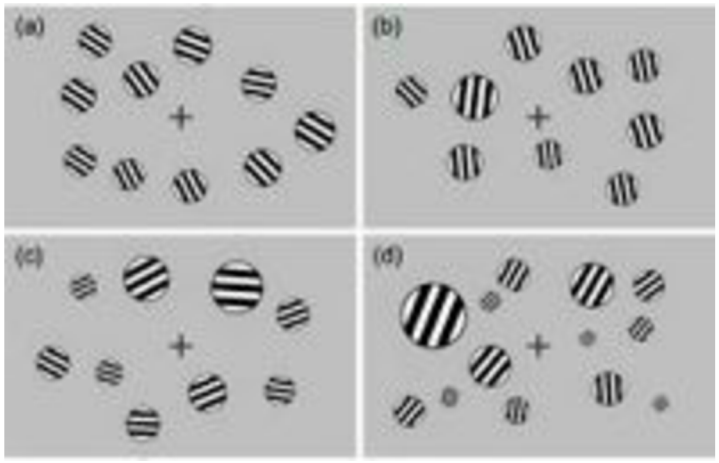


Figure 2. 3 Cartoon examples of arrays from Ensemble block, with different levels of external noise- (a) 1°, (b) 3°, (c) 6°, and (d) 10°

I ensured that the individual sizes of gratings in each display adhered to the specified mean and SD for each trial, with tolerances of $\pm$0.16° (i.e., four pixels) for the mean and $\pm$0.02° for the SD. Though this restriction violated the assumptions of random sampling, the deviation from random sampling was small given the set sizes that I used, and this restriction ensured that the trials within each block accurately reflected the target mean and SD while still allowing for sufficient variability across trials to fit the variance summation model. To provide an accurate fit to the data, I ensured that all analyses took into account the actual average size of the individual elements presented to the subjects, instead of the values suggested by QUEST. Since the average orientation of the gratings was not a relevant feature, the average orientation of the gratings for each display varied

randomly among six values (30˚, 60˚, 90˚, 120˚, 150˚, and 180˚), which allowed the gratings to look more distinct.

In the Single-grating block, I measured size discrimination thresholds for a single grating at the standard size of 2.8˚ in diameter. We included two different sub-blocks that varied the locations of the single gratings, one in a foveal region (4˚ x 4˚ around the fixation cross) and one in the periphery (from 56˚ x 40˚ of visual angle, but never appearing within the 4˚ x 4˚ foveal region). Including two blocks allowed us to estimate the reduction of size sensitivity as a function of eccentricity.

Each of the six sub-blocks (four sub-blocks for the Ensemble block and two sub-blocks for the Single-grating block) lasted approximately 15 min, and the order of these blocks was randomized across subjects. All blocks were run during a single session.

2.3.2 Modeling: Variance summation model

The variance summation approach enables one to measure how observers' response variability may change as a function of stimulus variability. Intuitively, it should be easier for an observer to estimate the average size of an ensemble when variability in size is low, and performance should become poorer as variability increases (Figure 2.3) The variance summation approach exploits a noise analysis that assumes the additivity of variances on the basis of convolution to model the data (Equation 1) in terms of the local and global limits of the system and external noise. In the variance summation model, the local and global limits are characterized by the internal noise of the ensemble averaging mechanism and the sample size that the observer gathers from the stimulus, and the external noise is assessed by the variability embedded within the stimulus, such that

$$\sigma_{obs} = \sqrt{\sigma_{int}^2 + \sigma_{ext}^2/n}\,, \qquad [1]$$

where σ_{obs} is the observed threshold, σ_{int} is the intrinsic or internal noise, σ_{ext} is the external noise, and n is the number of samples being employed. In the experiments of Chapter 2, σ_{ext} is the variability of the sizes within an ensemble of sine gratings (e.g., Figure 2.3), which is under the experimenter's control (i.e., the Gaussian distribution of object sizes in the display). σ_{int} is the noise or error inside the head of the observer (also assumed to be Gaussian) that affects their estimate of the ensemble average. Thus, Equation 1 is simply a way of combining these two Gaussian sources of noise in order to fit the observed sensitivity of the observer.

By measuring observed thresholds (σ_{obs}) at multiple levels of external noise (σ_{ext}), it is possible to fit values for the internal noise affecting the ensemble averaging mechanism (σ_{int}) and the number of samples (n) that the observer seems to rely on (i.e., the number of individual gratings a subject averages). When the external variability (σ_{ext}) is lower than the internal noise (σ_{int}), the observed threshold (σ_{obs}) will derive almost entirely from the internal noise. But as the external variability (σ_{ext}) increases, it will eventually come to exceed the internal noise (σ_{int}) to become the dominant force determining the observed threshold (σ_{obs}). Intuitively, the observed threshold will not increase rapidly until the external noise is greater than the internal noise.

Sample size will function to raise or lower the observed thresholds (σ_{obs}) across all levels of external noise (σ_{ext}), as pooling evidence from greater numbers of items will result in reduced observed thresholds (σ_{obs}). The pattern of reduction in the observed thresholds due to increased sample size is distinct and separable from the reduction that occurs from reduced internal noise. The approach used in Chapter 2 was inspired by

previous research in which variance summation modeling has been used to estimate the internal noise and efficiency (i.e., sample size) of texture discrimination mechanisms (Beaudot & Mullen, 2005; S. C. Dakin, 2001; Steven C Dakin, Bex, Cass, & Watt, 2009; Demanins, Hess, Williams, & Keeble, 1999; Heeley, Buchanan-Smith, Cromwell, & Wright, 1997). A benefit of this approach is that, once generated, estimates of the internal noise affecting ensemble processing can be compared to behavioral estimates of the internal noise affecting individual object processing to address the question of whether or not the internal noise for ensemble processing is lower than that for processing individual items.

2.3.3 Results and Discussion

In the Ensemble blocks, we systematically manipulated the variability of the sizes within each ensemble as a source of external noise (Figure 2.3). If ensemble representations pool evidence across items that vary in size, the observed thresholds should increase systematically with increasing external noise. Figure 2.4 displays the observed thresholds (σ_{obs}). The thresholds smoothly increased as external variability (σ_{ext}) increased.

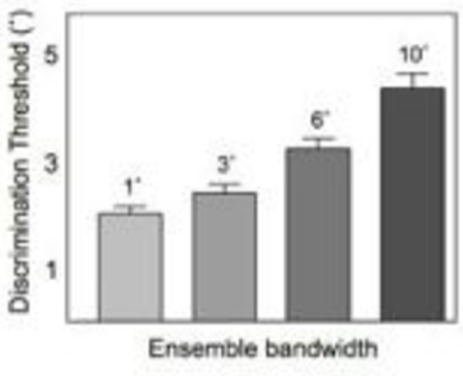


Figure 2. 4 Results: Discrimination thresholds for the ensemble conditions at each bandwidth

In an ensemble process that pools evidence across multiple samples, the specific pattern of increase in the observed thresholds as a function of increasing external noise should be systematically related to a relationship between sample size (n) and the internal (σ_{int}) and external (σ_{ext}) noise. This relationship can be formalized by the additivity of variances, as in Equation 1 (Beaudot & Mullen, 2005; Dakin, 2001; Dakin et al., 2009; Demanins et al., 1999; Heeley et al., 1997). This approach has been used successfully to estimate the internal noise and sample size for average orientation processing (Beaudot & Mullen, 2005; Dakin, 2001; Heeley et al., 1997). In this model, the manner in which the average-size thresholds (σ_{obs}) increase as the external variability in sizes increases (σ_{ext}) can be determined by a summation of noise processes.

I fit the data from the ensemble size blocks for each subject separately using the variance summation model (Equation 1) to obtain estimates of the internal noise and the number of samples involved in the averaging process using least-squares estimation. The group fit can be seen in Figure 2.5, where the observed thresholds from Figure 2.4 are reprinted as data points and the model fit is a smooth curve. The model provided an accurate fit to the subjects' performance as a function of bandwidth ($R^2 = .99$, $p < .01$).

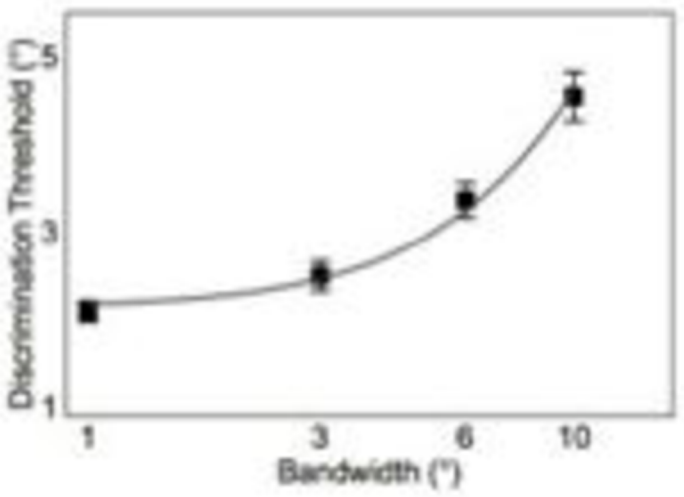
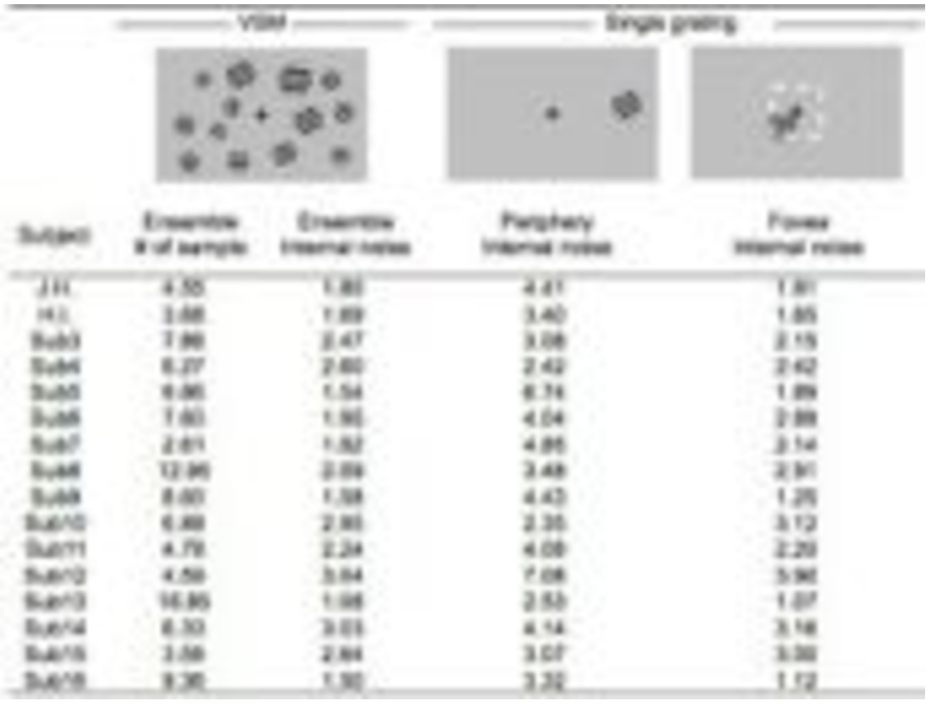


Figure 2. 5 Results and model

One of my central interests was to use the variance summation model to address Myczek and Simon's (2008) suggestion that only one or a few individual objects need to be sampled in order to attain the performance of human subjects in average-size discrimination tasks. The group fit from the variance summation model in Figure 2.5 determined the estimate of 7.0 samples from each display. Figure 2.6 presents the number of samples and the internal noise determined for each subject. An estimate of 7.0 samples per display exceeds the widely discussed (but not uncontroversial) estimate of a three- to four-item object-based limit of parallel attention (Oksama & Hyönä, 2004; Z. W. Pylyshyn & Storm, 1988; Scholl, 2001). This analysis suggests that Myczek and Simons underestimated the number of samples that subjects rely on in ensemble feature tasks of average-size processing. One reason for their lower estimate may be that their simulations did not take internal noise (σ_{int}) into account (Ariely, 2008; Haberman &

Whitney, 2010). Estimating the internal noise that affects the processing of ensemble average size was my second major focus.



Subject	Ensemble # of sample	Ensemble Internal noise	Periphery Internal noise	Fovea Internal noise
J H	4.55	1.80	4.41	1.91
H L	3.68	1.89	3.40	1.85
Sub3	7.86	2.47	3.08	2.15
Sub4	6.27	2.60	2.42	2.42
Sub5	6.86	1.54	6.74	1.89
Sub6	7.60	1.90	4.04	2.88
Sub7	2.61	1.52	4.85	2.14
Sub8	12.96	2.09	3.48	2.91
Sub9	8.00	1.58	4.43	1.25
Sub10	6.89	2.90	2.35	3.12
Sub11	4.79	2.24	4.09	2.29
Sub12	4.59	3.64	7.08	3.96
Sub13	16.90	1.98	2.53	1.07
Sub14	6.33	3.03	4.14	3.16
Sub15	3.58	2.84	3.07	3.00
Sub16	9.36	1.90	3.32	1.72

Figure 2. 6 Fitted parameters from the variance summation model and the internal noise for individual object blocks

If the representation of ensemble features relies on sampling individual objects and then averaging them- and perhaps throwing away the original samples (Alvarez, 2011; Ariely, 2001) - then the internal noise (σ_{int}) estimated by the variance summation model should match the observed thresholds for processing individual items. If, instead, the ensemble pooling process relies on estimating scene statistics without individuating items (e.g., perhaps through mechanisms similar or identical to texture processing), then the internal noise that affects this process may be distinct from the noise affecting individual object representations. There has been some suggestion in the literature that the internal noise affecting ensemble averaging may be lower than the internal noise

affecting representations of individual items (Alvarez, 2011; Alvarez & Oliva, 2008; Ariely, 2001; Chong & Treisman, 2003; Haberman & Whitney, 2009). Such suggestions were motivated by evidence that the observed thresholds in an ensemble averaging task tended to be lower than the observed thresholds for identifying individual items. But, observed thresholds can be lower due either to decreased internal noise or to increased numbers of samples, and the previous work was unable to disentangle the potential contributions of these two sources. The variance summation model allows us some handle on this question, as it allows us to measure the contributions of each of these factors from performance within the ensemble task itself.

We determined subjects' thresholds for processing individual gratings within both the fovea region and the periphery. Internal noise for ensemble processing across the entire display (Figure 2.6) was significantly lower than the observed thresholds for discriminating single gratings in the periphery ($t(15) = 4.77$, $p < .01$) and marginally lower than the thresholds for discriminating single gratings in the fovea ($t(15) = 1.59$, $p = .13$; Figure 2.6). Importantly, given the crowding controls in our ensemble displays, only one or two gratings could appear within the fovea region in ensemble displays- the remaining gratings would have appeared in the peripheral region. Because subjects relied on many more than one or two gratings during the ensemble feature task, in order to approach the performance of the human subjects, any object-based sampling strategy would need to assume fovea-level noise across all sampled items, not just for the one or two that happened to fall within the fovea region (or, it would need to sample even more items than was suggested by the variance summation model; Figure 2.6). Because the internal noise for representing individual gratings increases as one moves into the

periphery (e.g., note the differences in the observed thresholds for the fovea and periphery in Figure 2.6), it would appear that subjects do not rely on selecting and averaging individual gratings during the ensemble feature task.

Thus, variance summation modeling of performance in an average-size ensemble feature task suggests that the number of samples required for ensemble processing is greater than one or two items and that the internal noise affecting ensemble processing is lower than the internal noise for processing a single item presented in the periphery, and marginally smaller than the internal noise for processing a single item presented within the fovea.

In Chapter 2, I systematically manipulated the external variability of sizes within a set of sine gratings in an ensemble feature task in order to empower variance summation modeling to estimate the sample size (n) and internal noise (σ_{int}) that affect subjects' processing of average-size information. I also compared these estimates to the observed thresholds for processing sine gratings. I found that subjects appear to rely on many more than one or two individual gratings when representing the average size of items in an ensemble (Figure 2.6). I also found that the internal noise affecting the ensemble process is slightly lower than the internal noise that affects the representation of individual item sizes within the fovea. The variance summation modeling provides a means for studying internal noise and sampling procedures on the basis of performance within the ensemble feature task itself, and the results from the variance summation model suggest that ensemble processing relies on a mechanism that is distinct from the processing of single items.

CHAPTER 3. ENSEMBLES IN PERCEPTION, 2: PRECISION AND BIASES FOR THE ENSEMBLE FEATURE APPROXIMATE NUMBER ARE AFFECTED BY VISUAL GROUPING

3.1 SYNOPSIS

Perceptual grouping is known to be a rapid, pre-attentive process that gives rise to "higher-units" of representation as generated by an interpretation of configurations in an image (for review, see Chapter 3.2.1). In Chapter 3, I aim to build a connection between this rapid global processing for perceptual groups and ensemble representations. First, I introduce a new modeling approach to determining a formal description of the perceptual grouping behavior of human observers (Chapter 3.3). This approach relies on the K-means clustering algorithm from computer vision (detailed description is provided in Chapter 3.3.2). In Section 3.3, I show that this K-means clustering algorithm can provide a very accurate fit to human subjects' estimates of the number of clusters in random dot arrays. Then, in Section 3.4, I show that subjects tend to underestimate the number of dots in a display when the image contains many clusters (as fitted and predicted by the clustering algorithm). In the conclusions, I suggest that perceptual groups may serve as units for representation of approximate numbers, allowing for rapid extraction of ensemble features from briefly flashed visual scenes.

3.2 BACKGROUND

3.2.1 Perceptual grouping

One possible candidate for the mechanism supporting the rapid extraction of ensemble representations may be perceptual grouping and mechanisms that work over perceptual groups. Humans can readily and near-instantaneously organize the global structure from a visual scene by grouping multiple items together. Visual grouping has been a significant focus of perception research since it was first emphasized by Gestalt psychologists (Kubovy & Podgorny, 1981). The Gestalt psychologists argued that the visual system does not simply collect and combine sensory information from the external world to form a picture of the world, but instead actively organizes it using various laws of grouping, such as proximity. The law of grouping by proximity states that "when the [stimulus] field contains a number of equal parts, those among them which are greater in proximity will be organized into a higher unit" which "must be considered as real as the organization of a homogeneous spot" (Koffka, 1935, pp. 164-165). This "higher unit" refers to the product of a perceptual process that actively imposes structures upon the incoming sensory information. According to Gestaltists, this "higher-unit" of representation is an interpretation of configurations in an image, and it cannot be derived simply by examining any constituent parts in the image in isolation. The mental computations that are dictated by the law of proximal grouping have been suggested to be computed in a bottom-up fashion using relatively local information by algorithms that are purely data-driven (Pomerantz, 1983) and achieved at a preattentive stage of the visual-processing hierarchy (Neisser, 1967).

The product from this grouping process then provides the inputs to later processing states (Palmer & Rock, 1994; Vecera & O'Reilly, 1998). For example, it has been explicitly noted by earlier studies that how elements are grouped in an image affects how one perceives the visual number of individual elements in the image (Woodworth & Schlosberg, 1954). Depending on spatial configurations the same number of dots may result in vastly different experiences of the apparent number of their elements. For example, items that are separated further from one another, together occupying a larger area on the display, are usually perceived to be more numerous (Bevan, Maier, & Helson, 1963; Krueger, 1984). Differently located items in the stimulus appear more or less numerous depending on how they are distributed such that globally-clustered items, e.g., a homogenous density of items, appear to be more numerous than the same number of items clustered into multiple sub-groups (Frith & Frith, 1972). Regularly-arranged items look more numerous than randomly-distributed items (Ginsburg, 1976; Taves, 1941) and random patterns look more numerous than clustered items (Ginsburg & Goldstein, 1987).

In addition, the grouping pattern in a display also affects number-estimation latencies such that dots that are randomly spread out in the periphery of the display are enumerated faster than the same number of dots tightly clustered in the center of the display (van Oeffelen & Vos, 1982).

Such biases in estimation can be understood to emerge from Gestalt's grouping principles in which parts (i.e., the individual dots), which join to form a good figure (e.g., a regular shaped homogenous cluster in the center of the display), will be experienced as less separate than items which join to form multiple clusters or an ill-formed cluster (e.g., an irregularly shaped-cluster or two clusters that are spread out into the periphery). That

is, a good configuration of items may be grouped into a single unit and tend to serve "as real as the organization of a homogeneous spot (Koffka, 1935)".

When considering the grouping of many items into sets or chunks, it is important to consider the cognitive and processing limitations involved in human set representation. Just as with representations of individual objects, set representations of perceptual groups are constrained by visual working memory (Nelson Cowan, Chen, & Rouder, 2004; Halberda & Feigenson, 2008). The number of sets that can be represented never seems to exceed the capacity limits of visual working memory (three or four set limit in VWM; Halberda & Feigenson, 2008). Moreover, sets that are formed from multiple individual items can further be bound into a "super-set", and – when a super set is formed by grouping separate subgroups together – the number of sets that can be bound into this super set also seems to obey these memory limits (Chase & Ericsson, 1981).

Given the limited capacity of VWM, this set-based representation by grouping could increase the amount of information that can be maintained in memory because multiple individual items can be grouped into a set and stored together in memory. For example, Xu and Chun (2007) have shown that perceptual grouping enhances visual working memory by allowing more visual elements to be remembered when they were grouped. Woodman, Vecera, and Luck (2003) have also shown that perceptual grouping influences what elements are stored in memory such that when one element of a group was stored in working memory, other elements of the sample gestalt group were likely to be stored as well. These results suggest that set-based representation can be another unit for processing that functions like a single individual object for visual working memory.

How is this grouping process achieved? Atkinson, Cambell, and Francis (1976) related the grouping of information to the bandwidth of channels tuned to a particular size or spatial frequency. Within a region of a certain size, elements are less likely to be discriminated and segmented, giving a compulsory perception of a single higher-order object. If such channels were arranged hierarchically such that individuals were represented by finer channels and entire groups by coarser channels, then ensemble representation might be understood to be a statistical description of the activity in finer channels within a single coarser channel (e.g., average orientation from multiple crowded gabor patched; Parkes et al., 2001). In Experiments B and C, I investigate the possibility that perceptual grouping, or clustering, within random dot arrays will affect human estimates of the ensemble feature approximate number.

Although the validity of perceptual grouping seems intuitively apparent, formal descriptions of the underlying mechanisms have been lacking: much evidence initially presented in support of perceptual grouping has been limited to phenomenological demonstrations. This is surprising given that grouping by proximity is one of the most well known and intuitively appealing principles. Only a few attempts have been made to propose and evaluate formal models of perceptual grouping (e.g., CODE algorithm proposed by van Oeffelen & Vos (1982) and evaluated by Compton & Logan (1993)). And, existing computational models of perceptual grouping might be improved and refined such that the models would be capable of explaining human grouping behavior more efficiently, with simpler estimation procedures (e.g., with fewer number of free parameters). Here, I propose and test one such approach.

In Experiment B, I describe the K-means clustering algorithm, which I will adapt from computer vision research, and I present evidence that it provides an adequate fit to human observers' judgments of the number of clusters within random dot arrays. This algorithm has 1 free parameter (e.g., compared with the 5 free parameters of CODE, van Oeffelen & Vos, 1982). In addition, in Experiment C, I find evidence that the clustering estimates generated by K-means clustering provide a viable mechanism for the well known phenomenon of human underestimation in enumeration tasks. To wit, the more "clustery" an array is – as judged by K-means clustering – the more human observers tend to underestimate the total number of items in the array.

Perceptual groups can be organized at different levels of perceptual hierarchy (e.g., lower levels of hierarchy encoding the more specific details of an image vs. higher levels organizing the details into more global structural units; Palmer, 1975). Figure 3.1 illustrates the idea. In Figure 3.1a, perhaps we see only one cluster of black dots. But, other grouping possibilities can be highlighted using color. For instance, an observer might experience two groups in Figure 3.1a, as diagrammed using colors in Figure 3.1b - one group of 10 dots and another with 7. Or, another observer might experience seven groups of 2 or 3 each (as in Figure 3.1c). In this way, grouping by proximity can operate under a "tight" grouping criteria - with only the closest elements being grouped (e.g., seeing seven groups of dots; Figure 3.1c) – or under a "loose" grouping criteria (e.g., seeing two groups of dots; Figure 3.1b). .

In Section 3.3.2, I introduce a new approach to modeling perceptual grouping judgments of human observers in images containing randomly located dots. For simplicity, the current model contains one free parameter and utilizes an efficient

optimization algorithm based on K-means clustering. The free parameter of the current model reflects the size of the grouping window for an individual subject, determined independently for each of the stimulus images. The size of grouping window determines the grouping strength, with smaller grouping windows resulting in less items being grouped together (i.e., "tight" grouping criteria), and with larger grouping windows resulting in more items being grouped together (i.e., "loose" grouping criteria). I demonstrate that this K-means clustering model can provide a formal description of human grouping judgments in an efficient and flexible manner with a single free parameter.

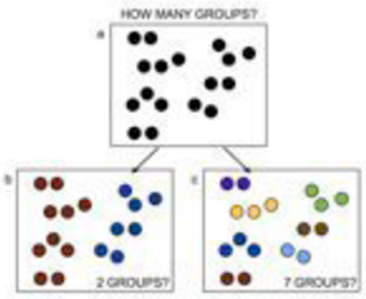


Figure 3.1 Different possibilities of grouping

3.3 EXPERIMENT B: ASSESSING OBSERVERS' ABILITY TO ESTIMATE THE NUMBER OF CLUSTERS WITHIN AN ENSEMBLE

3.3.1 Experimental method

Subjects

10 naive undergraduate students participated in the experiments for course credits. All of the subjects had normal or corrected-to-normal vision.

Apparatus and stimuli

The stimuli were generated using MATLAB software, together with the Psychophysics Toolbox extensions (Brainard, 1997; Pelli, 1997), and were displayed on an LCD monitor driven by a Macintosh iMac computer (the viewable area was a gray central square window with a 17-in. diagonal). The subjects were seated approximately 50 cm from the screen and viewed the display binocularly. At this viewing distance, each pixel was approximately 0.04 of visual angle. The stimuli were presented on a gray background and consisted of multiple blue dots (5-35 dots) each of which subtended 0.96˚ of visual angle. Location of the dots were randomly chosen for each of 180 visual images, within the invisible gray area, subtending 16˚ x 20˚ of visual angle.

Procedure

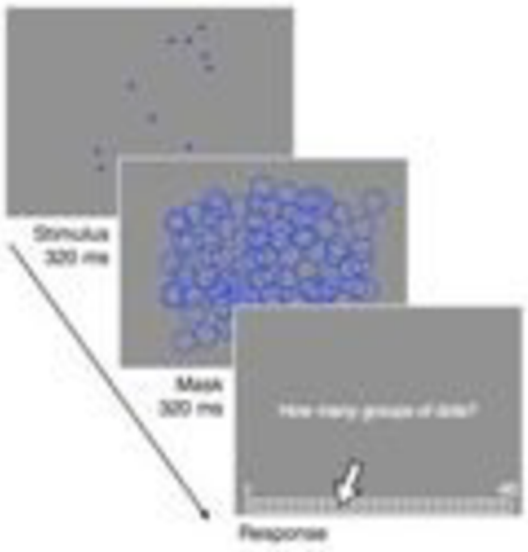


Figure 3. 2 A sample trial of Experiment B

Figure 3.2 illustrates a sample trial of the experiment. All 10 subjects were presented with the same 180 stimulus images that were generated in advance, but with a different sequence of the images. After a ready signal, the stimulus array containing multiple dots was presented for varying durations (from 50 msec to 320 msec), followed

by a mask array and response array. For the response array, there appeared a linear response scale with continuous values from 1 to 40. Subjects were instructed to click on any value between 1 and 40 to make a response, using a mouse cursor.

The task instruction given to the subjects for Experiment B was minimal: the subjects were asked to simply judge how many groups of dots were presented. The subjects were told that there was no right / wrong answer, that they did not need to group items at all if they thought that none of the elements seem to be grouped (if this was the case the subjects could simply report the number of individuals) and that they could group items however they wanted and felt the most comfortable and natural way.

3.3.2 Modeling: A computer vision approach for modeling grouping within ensembles

In order to systematically assess the grouping present in each visual image, I applied a K-means clustering algorithm – one of the popular clustering analysis techniques that are used in computer vision literature. In computer vision, one of the common problems to deal with is segmenting input images so that the computer can do further image processing, such as scene and object recognition or image categorization. Image segmentation in computer vision is the process of partitioning an image into multiple segments (e.g., groups of pixels). The goal of segmentation is to simplify and change the representation of an image into something that is more meaningful and easier to analyze. In the K-means clustering algorithm, the machine iteratively partitions an image into K clusters in order to settle on a final output that appears to be the most reasonable segmentation of the image, given the constraints of the algorithm. The basic algorithm is as following (see also Figure 3.3):

1) The computer picks K initial "means" (In Figure 3.3, three means are randomly generated within the data domain).

2) K clusters are created by associating every observation (e.g., each dot) with the nearest mean.

3) The centroid of each of the K clusters becomes the new mean.

4) Steps 2) and 3) are repeated until convergence has been reached (e.g., no pixels change clusters).

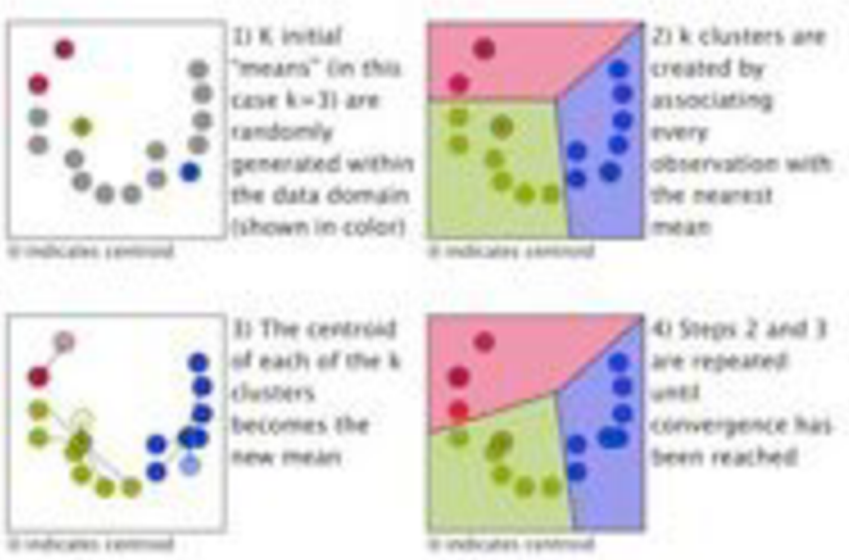


Figure 3. 3 Basic algorithm of K-clustering

Based on the basic algorithm of K-clustering in computer vision, I utilized a centroid-based clustering algorithm in which the model aims to partition the N_d dots into k sets ($k \le N_d$), so as to minimize the number of clusters k with the constraint in Equation [2]:

$$\left\| x_p - m_i \right\| \le T_d, \forall x_p \in S_i, \qquad [2]$$

where x_p is the location of each of the dots that belong to a given cluster i in an image, m_i is the center location of the cluster, T_d is the threshold distance that can be assigned to the same cluster, and S_i indicates the index of each cluster i. Similar to the conventional algorithm for a k-means clustering (Lloyd, 1982), the algorithm proceeds by alternating between two steps:

1) Assignment step: Assign each observation to the cluster with the closest mean:

$$S_i^{(t)} = \{x_p : \left\| x_p - m_i^{(t)} \right\| < \left\| x_p - m_j^{(t)} \right\|, \forall 1 \le j \le k^{(t)}\} \qquad [3]$$

2) Update step: Calculate the new means to be the centroid of the observations in the cluster.

$$\mathrm{m}_i^{(t+1)} = \frac{1}{\left|S_i^{(t)}\right|} \sum_{\mathrm{x}_j \in S_i^{(t)}} \mathrm{x}_j \qquad [4]$$

Critically T_d which is the clustering threshold, defined as the center-to-center distance in which elements can be grouped together was set to be a free parameter. In other words, the clustering threshold serves such as a window size for grouping such that if elements are closer to each other than the size of this grouping window with the diameter of the clustering threshold, they will be grouped together. If elements are further than the size of the grouping window, however, they will not be grouped. Therefore, T_d determines the extent to which elements are grouped together. For example, if T_d is large, more and more items will be grouped together while if T_d is small, fewer and fewer items will be grouped. This is a novel approach that has not previously been taken in the computer vision literature. This approach allows me to use T_d - and the K-means clustering algorithm - as a formal specification of what might be described as the human

visual system's maximal grouping distance. By fitting the algorithm to the same images that human subject saw and determining the number of clusters the model settles on for a range of T_d, this approach will allow me to determine what value for maximal grouping distance (T_d) best matches the judgments of human subjects.

The model runs the three-step routines iteratively at varying T_d values and the free parameter T_d was determined to be the value that yields the minimum deviance from the actual human observer's response on the number of groups for the stimulus images. The model fits the free parameter T_d for each subject and for each stimulus image, allowing for comparison in grouping window sizes across human subjects in a formally constrained and quantitative manner.

3.3.3 Results and Discussion

It is worth emphasizing that in Experiment B, every single response for each visual stimulus is a subjective measure, therefore there are no correct or incorrect responses: subjects could group items in whatever way they felt the most natural and comfortable. They were free to report how many groups they thought they would parse from a given image. Hence, the first analysis for Experiment B is rather qualitative and comparative. Specifically, I first compared responses from the 10 subjects on each of the images and asked to what extent their responses agreed with one another. Figure 3.4 shows a few examples of the images shown to the subjects and histograms of the 10 subjects' responses for each of the images. Even though there was no specific instructions

for how to group items, the subjects provided similar responses for the number of groups in each image, suggesting that all observers found this to be a natural and intuitive task.

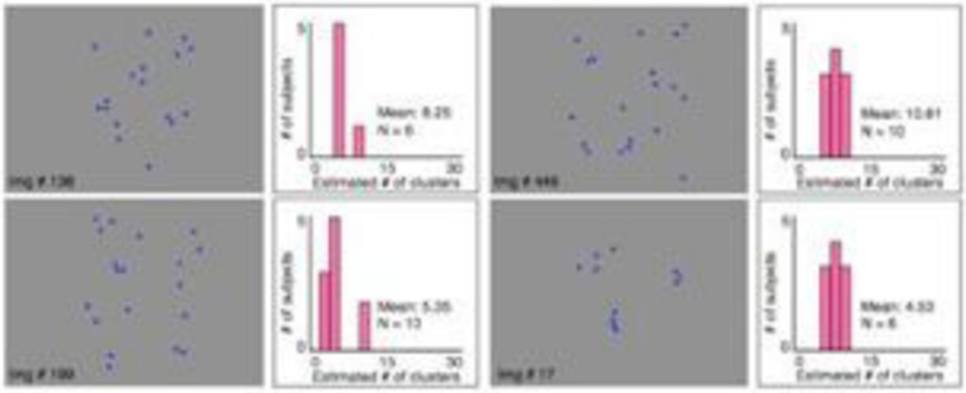


Figure 3. 4 Results: example of responses on the four selected stimulus images

Figure 3.5 shows scatter plots for the behavioral responses from three individual subjects. In each plot, each circle represents a particular image from the180 image stimulus set, with the x-value being the behavioral response from one subject and the y-value being the behavioral response from another subject. In this way, the scatter plots in Figure 3.5 display the extent to which the two subjects agreed in their estimates. The R^2 values reveal quite robust agreement in behavioral responses across subjects. Despite the open-ended nature of the task, and the fact that the subjects could choose their own criterion for how to group items, the grouping pattern was highly consistent across individuals.

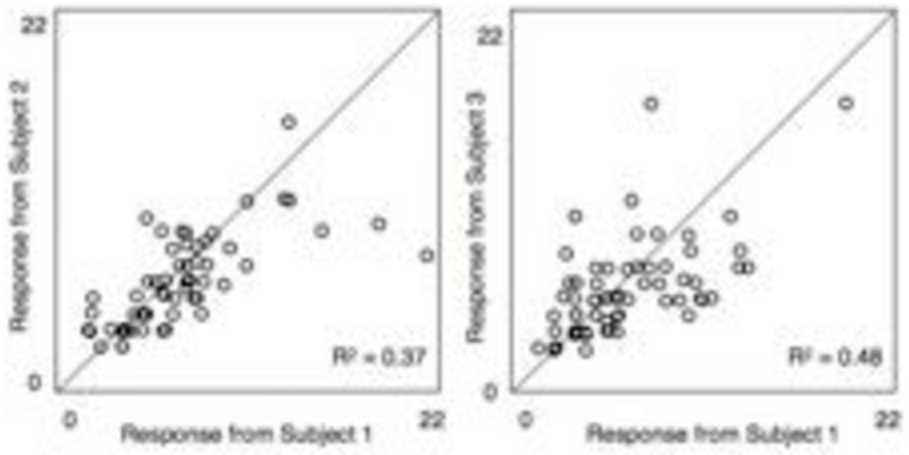


Figure 3. 5 Examples of correlation between responses from different subjects

The clustering model with a single free parameter for center-to-center distance among items (i.e., clustering threshold, T_d) was fitted to the human subjects' responses and provided the best-fit clustering threshold value for each image and each subject. The best-fit clustering threshold was determined at each stimulus duration by calculating the deviance between the model prediction and the human estimation on each of the stimulus images at varying distance thresholds, for each individual subject. Figure 3.6 first shows a histogram of all the best-fit clustering thresholds from all the presented stimulus images and from all the subjects. The majority of the best-fit clustering threshold values were around 4° of visual angle as a diameter of the grouping window.

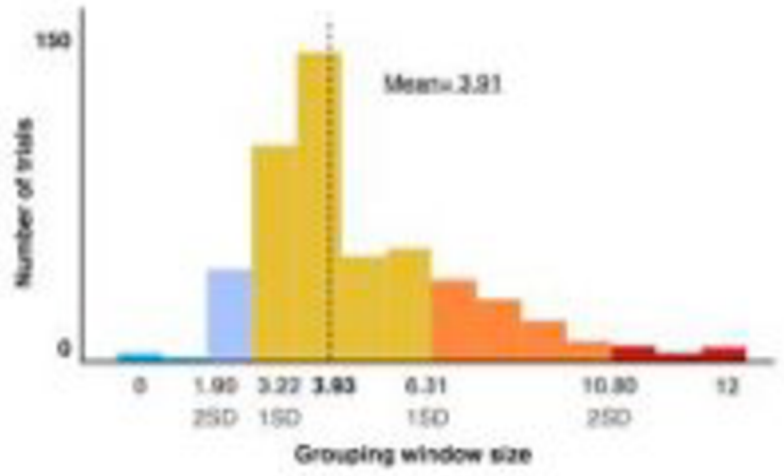


Figure 3. 6 Histogram of the best-fit clustering thresholds

In addition, Figure 3.7 plots the average of the best-fit parameters for the clustering thresholds at different stimulus durations (from 50 msec to 330 msec). The best-fit clustering threshold values did not change over time, suggesting that clustering of items into groups occurred as fast as 50 msec, consistent with the previous notion that perceptual grouping is a fast, pre-attentive process.

To assess the stability of the model estimates for different stimulus images presented throughout the trials within subjects, I averaged clustering threshold values obtained from the 180 stimulus images for each subject. The average clustering threshold value was used when the model predicted individual subject's response on each stimulus image. By using the average of fitted clustering threshold values, the model could reliably predict each subject's responses to the images. Figure 3.8 shows the scatter plot of the model-predicted number of clusters assuming one fixed value of average clustering threshold for each subject and the subject's response to the stimulus images. The overall R^2 value reveals that the model prediction based on the averaged estimates of clustering

threshold well captured the actual subject's estimation for each image (the average of R^2 = 0.623), suggesting the model estimates were stable and consistent within subjects.

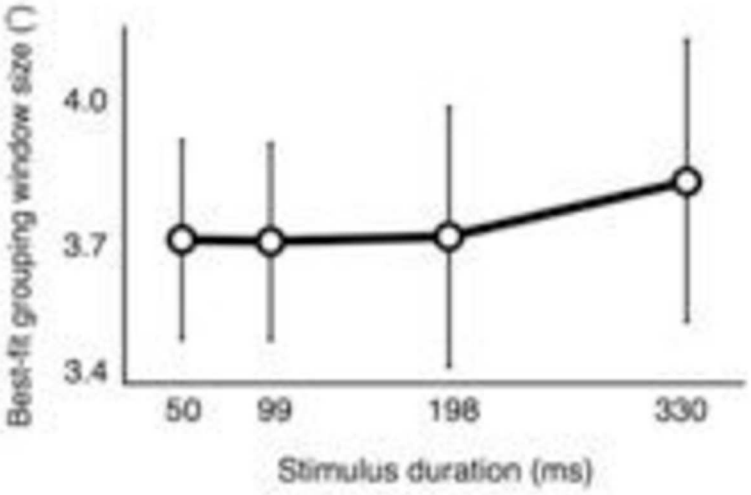


Figure 3. 7 Average of the best-fit grouping window size at different stimulus durations

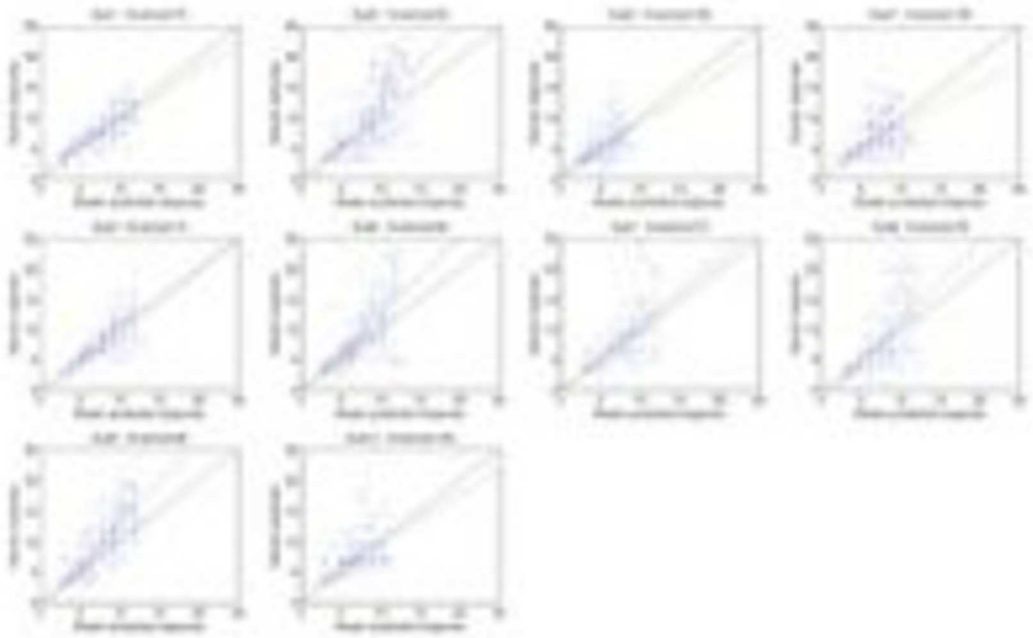


Figure 3. 8 Model prediction of each human subject using the best-fit grouping window size

Together, these results suggest that human subjects' grouping pattern highly agreed with one another and the subjects' estimation of the number of clusters of elements in stimulus images was well captured by a simple model with one free parameter for the grouping window size. The perceptual grouping in the stimulus images containing randomly positioned dots happened very fast, and possibly in a pre-attentive manner, as fast as in 50 msec. To my knowledge, the current study is the first attempt to compare quantitatively and systematically the human subjects' impression of perceptual groups in visual images. The subjects' impression of perceptual groups seems to be a "common sense", meaning that human subjects highly agree with one another in organizing perceptual groups and estimating the number of perceptual groups, despite that visual input from the images does not favor any one specific grouping strategy. It seems that human subjects tend to exploit similar rules for determining whether items are grouped together or not, by using similar grouping threshold which is defined here as the maximum center-to-center distance among elements in a perceptual group.

The proposed clustering algorithm revealed that the critical distance for determining whether elements are grouped into one cluster was approximately 4˚ of visual angle, as the diameter of grouping window. The critical distance for grouping was also consistent across individual subjects and across the stimulus images regardless of the actual number of items in each of the images. In different contexts, there have been interesting findings of 3.5˚- 4˚ of visual angle that may be of relevance to the current findings. For instance, the similar numbers, 3.5˚-4.0˚ were found to be a critical distance that differentiated qualitatively the patterns of human performance on multiple object tracking (MOT) tasks. Alvarez and Franconeri (2007) showed that when the minimum

distance between items was closer than 3.5˚ - 4.0˚, human subjects' accuracy in tracking moving target items was not impaired by increasing speed of the moving items. However, they showed that the subjects' accuracy was not affected by the speed of the items, yielding the comparable accuracy both for slowly-moving items and fast-moving items when the minimum distance between items was greater than 3.5˚-4.0˚. Similarly, this specific number of 3.5˚-4.0˚ of visual angle was also observed to be critical for the step function of human performance such that confusion non-target items happened when the items were closer than this distance of 3.5˚ and 4.0˚ whereas confusion did not happen when the items were further than this distance (Bae & Flombaum, 2012). Despite these interesting parallels in different contexts of visual processing, the best-fit grouping threshold of 4˚ of visual angle still requires further investigation in order for any claims about the theoretical implications of this specific number I discovered. For example, it should be further examined whether this best-fit grouping threshold of 4˚ visual angle is invariant to other factors of images, such as the scale, density, size of items, the total area of the visual field and so on.

3.4 EXPERIMENT C: CLUSTERING AFFECTS OBSERVERS' PRECISION AND BIASES WHEN EXTRACTING APPROXIMATE NUMBER

3.4.1 Experimental method

Subjects

In Experiment C, different 10 naive subjects participated in the experiments for course credits.

Apparatus and stimuli

All the aspects were identical to those in Experiment B.

Procedure

Subjects were presented with exactly the same 180 stimulus images that were used in Experiment B. The order of presentation of each image was randomized for each individual subject, as in Experiment B. On each trial, a stimulus array containing multiple dots was presented for 320 msec, followed by a mask array and response array. On the response array, a linear response scale with a range from 1 to 40 appeared, such that the subject clicked on any values between 1 and 40 to make a response, by using a mouse cursor. Unlike in Experiment B, the subjects were asked to estimate the number of individual dots, not the number of clusters. No feedback was given to the subjects.

3.4.2 Results and Discussion

If perceptual groups of elements in a stimulus image provided a unit for this rapid extraction of numerosity of individual dots, one would expect that the subjects' strong tendency to underestimate the number of individual dots could be well captured by the subjects' grouping pattern in each of the stimulus images. Therefore, I first asked how different grouping patterns in different images affected the subjects' estimation of the number of individual elements. I calculated the clusteredness index that indicated how many elements were clustered together in a stimulus image, based on human subjects' responses. The clusteredness index was calculated by dividing the actual number of dots

presented on an image which was reported by human subjects by the number of clusters that clustering algorithm predicted at the human subjects' best-fit grouping window size, at approximately 4° of visual angle. The larger clusteredness index, therefore, means that human subjects grouped more elements together (densely clustered) and the smaller clusteredness index means that the human subjects did not tend to perceive perceptual groups from the stimulus images, most likely because the spatial configuration of the dots in the image was not ideal for perceptual groups to be perceived (e.g., the dots were scattered enough, etc.).

According to the clusteredness index, each of the 180 images shown to the subjects was classified into one of the four categories with different degrees of clusteredness (Figure 3.9). The slopes of the regression lines in the plots of the number of dots and human responses were calculated separately for each of the four categories of clusteredness. The Figure 3.9 shows the average slopes for the four categories.

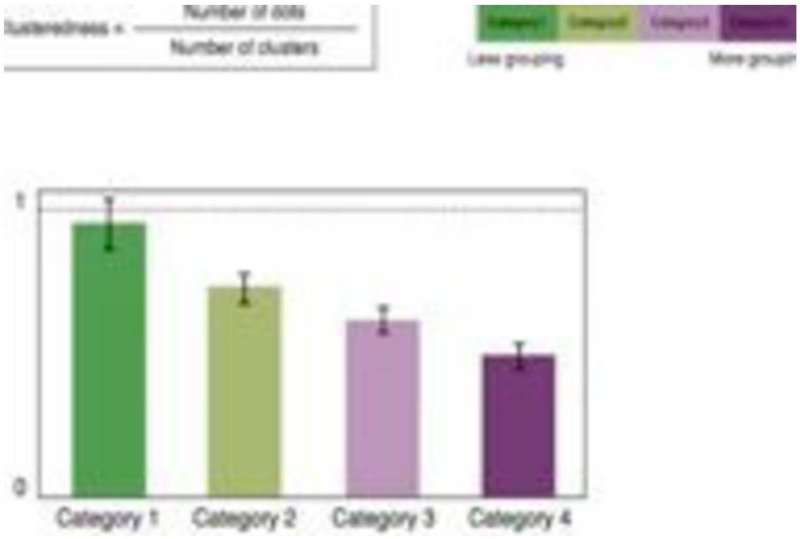


Figure 3. 9 Average slope for four categories of clusteredness

The slopes increased as the estimated clusteredness of the image increased and the one-way ANOVA revealed that this effect by the clusteredness category of images on the slopes was significant ($F(3,71) = 14.48$, $p < .01$). Additional one-sample t-tests revealed that the slopes for the categories 1-3 were significantly lower than 1 (Category 1: $t(17) = -11.76$, Category 2: $t(17) = -10.36$, Category 3: $t(17) = -5.96$; all p's $< .01$), indicating that the subjects' underestimation of numerosity of the dots for the three "more-clustered" categories, however the slopes for the image that belonged to the "least-clustered" category were not significantly different from 1 ($t(17) = -0.54$, $p = 0.60$), indicating no underestimation when the dots were not clustered. This result suggests that when human subjects tend to group more items together from a given image, they also tend to underestimate the number of elements in the image and vise versa. When dots are perceived to be more clustered and grouped on a display, they are perceived as being less numerous. However, when the dots are not perceptually clustered, they are not underestimated at all. These results suggest that the signature of numerosity representation could be well explained by human subjects' grouping pattern. In specific, when human subjects perceived the stimulus image to be more clustered with the impression of only a few perceptual groups, they also tended to more underestimate the number of individual dots of the image. On the other hand, when they perceived that the image was not clustered with the impression that there were more perceptual "groups", their tendency to underestimate the numerosity of the dots disappeared.

Taken together, these results support the idea that representation of approximate number of items may be possible at the level of perceptual groups rather than individuated objects, allowing for faster and more global processing for ensembles.

Perceived groups in a visual image have been considered as compulsory product from automatic, pre-attentive processes (Neisser, 1967). This efficient global representation built from the low-level scene structure can be also used for representing ensemble features. Such global process will allow the visual system to register multiple elements (possibly more items than those the system can process at a time) in a parallel manner, giving rise to rapid extraction of ensemble features. Perceptual grouping may serve as a form of a primitive chunking such we hierarchically reorganize items in the stimuli and group some of them together as a unit for further processing, thereby decreasing the amount of information we should process at once. In this manner, perceptual grouping allows for parsing multiple nested-levels of representation of the same stimuli from individuated items to one global scene representation. Halberda, Sires, and Feigenson (2006) suggested that in number estimation task, hierarchical coding of "group" and "individual" are both available for enumeration by the approximate number system. They further suggested that notion of a "group" may operate prior to enumeration of individuals by the approximate number system. The "groups" structured by perceptual grouping is a reasonable candidate for ensemble feature representation as well given the aspects of the visual images that are used for ensemble representation: many similar items are positioned randomly all over the display. It is therefore reasonable to expect that the elements are spatially grouped into clusters even before ensemble feature extraction. As previous researchers speculated, the "groups" built from this automatic grouping process is not easily overridden or split into single individuals, unless focused attention toward any single individual object is strongly forced to do so. Therefore,

perceptual groups built from the visual image may serve as a unit for ensemble representation, allowing for a fast, effortless process in global manner.

CHAPTER 4. ENSEMBLES IN ATTENTION: EACH ENSEMBLE GROUP FUNCTIONS AS A UNIT FOR SUBITIZING

4.1 SYNOPSIS

In Chapter 4, I argue that each ensemble group can be selected as a unit for attention. This is an important step in the processing of ensemble groups, because it links early perceptual processing of ensembles – discussed in Chapters 2 and 3 – to later cognition over ensembles (e.g., Working Memory) – discussed in Chapter 5. Here, I suggest that attentional selection based on an ensemble can maximize the efficiency of attention by saving on the demand of selecting and processing each element separately. In Experiment D, I rely on subitizing as an estimate of "unit"-based attentional selection. Here, I use the term unit-based in place of the more typical "object"-based, because the focus of Chapter 4 will be to demonstrate that each ensemble – consisting of multiple similar objects – can function as a single unit for visual indexing and can be subitized just as individual objects.

4.2 BACKGROUND

4.2.1 Units of selection in visual attention: Objects versus ensembles

Processing objects in a visual scene requires deployment of attentive selection (Rensink, O'Regan, & Clark, 1997; Simons & Levin, 1997; Wolfe & Bennett, 1997). The classic study (Egly, Driver, & Rafal, 1994) on object-based attention demonstrated that it was faster to detect a target when it was located on a cued object than when it was the

same distance away, but on an uncued object, suggesting that object is a unit of attentional selection. Further work suggests that different features (e.g., color, size, or orientation) that belong to a single object are integrated and bound together and selected as one single unit (Treisman & Gelade, 1980).

Visual indexing is the mechanism by which certain salient features or objects in a visual display are indexed (Pylyshyn, 1989). It has been suggested that this type of visual indexing occurs before, and guides, attentional selection. It may be that indexing is limited in addition to the limits of object-based attention. Intuitively, it seems unlikely that indexing a scene – particularly a complex one – is performed through a series of index assignments (references to items) to each individual unit in that scene. Consistent with this intuition, the number of visual indexes employed at once appears to be limited (Pylyshyn, 1989). Many researchers have suggested that our ability to attend individual objects is limited to approximately 3-4 objects at any one time (Alvarez & Franconeri, 2007; Z. Pylyshyn, 1989; Scholl & Pylyshyn, 1999). But, a question arises: if seeing relies on visual indexing, and if indexes are limited, then how is scene perception even possible? I believe that this suggests that the visual indexing procedures must be more complex than individual references to individual objects.

One proposal for expanding the diversity of the visual indexing procedures is the possibility that one can index via an ensemble representation that is built from multiple individual items. In the case of groups of similar items, it seems highly inefficient to index individual objects. Instead we might attend those items through just a single reference to the entire group as opposed to requiring the computational power of attending to each item in that group. Representing an ensemble as a single unit of similar

items can therefore increase the efficiency of attention. Ensembles can provide compressed information about the general features of a set of multiple objects, saving on the demand of selecting and processing each element separately. This may explain how representing multiple objects as an ensemble enhances visual cognition (Alvarez, 2011): an ensemble consisting of N items is assigned a single visual index, rather than N indices, thereby empowering a more efficient use of the limited indexing capacity of visual processing.

The proposal that an ensemble group can be selected as a single unit for "unit"-based attention has yet to be explored. There are, as yet, no published tests of this proposal. Here I explore the possibility that each ensemble must be selected and attended to, much like a single object, and that the number of ensembles that may be selected at any one time is limited just as it is for individual objects. I focused on a well-known process that requires attention to individual units: subitizing (Egeth, Leonard, & Palomares, 2008). Enumeration for a small set of items is fast and accurate, and is referred to as subitizing (Kaufman, Lord, Reese, & Volkmann, 1949). However, for sets of items larger than about 3 or 4, a person typically begins enumerating slowly through a process of verbal counting, and enumeration is thus slower and more error prone for larger numerosities (Trick & Pylyshyn, 1994). This dichotomy between subitizing and counting, observed in response times and accuracy of enumeration, has been well established (e.g., Oyama, Kikuchi, & Ichihara, 1981; Taves, 1941; Trick & Pylyshyn, 1994). It has been explained as resulting from distinctions in visual working memory, FINSTs (Pylyshyn, 1989; Sternberg, 1966), separable neural systems (Piazza, Mechelli, Butterworth, & Price, 2002), and pattern recognition (Mandler & Shebo, 1982). The

majority of these accounts agree that each object in an enumeration task must be indexed, especially for numerosities above 4.

In this chapter, I test whether enumerating groups of items is also possible (e.g., counting the number of groups rather than the number of individual objects). If one presumes that ensembles actually receive a single index and function the same way an object does, then the same dichotomy should also exist for enumeration of ensembles, or groups of items. In other words, the performance of enumeration response times for ensembles (i.e., counting each group) should degrade similarly for individual items (i.e., counting each dot). Thus, enumeration of ensembles should result in a relatively constant response time throughout the subitizing range – i.e., for set sizes between about 1 and 3 – and a linearly increasing response-time as a function of set size thereafter, representative of the counting range.

4.3 EXPERIMENT D: SUBITIZING ENSEMBLES

4.3.1 Experimental method

Subjects

17 naive subjects participated in the experiment for course credits. All of the subjects had normal or corrected-to-normal vision.

Apparatus and stimuli

The stimuli were generated using MATLAB software, together with the Psychophysics Toolbox extensions (Brainard, 1997; Pelli, 1997), and were displayed on an LCD monitor driven by a Macintosh iMac computer (the viewable area was a gray central square window with a 17-in. diagonal). The subjects were seated approximately

50 cm from the screen and viewed the display binocularly. At this viewing distance, each pixel was approximately 0.04 of visual angle. Figure 4.1 illustrates the two different types of stimuli presented on the trials for Ensemble subitizing and Object subitizing. A stimulus for Ensemble subitizing consisted of between 1 and 6 clusters, each containing between 5 and 15 individual dots of the same color. The total area of each cluster was controlled to prevent numerosity judgment on the basis of total area. Each ensemble set was designated by a shared color and general location. Within each cluster for Ensemble subitizing stimuli, the number of dots was randomly determined and each dots' position was randomly determined under a constrained maximum cluster spread, defined as the largest distance between any two dots within a cluster. A stimulus for Object subitizing contained between 1 and 6 single rings of different colors. Ring size was determined to encompass the max coordinates (x, y) of a hidden ensemble collection that was generated according to the parameters for the ensemble displays. This ensured that ring sizes and cluster areas were balanced across the Ensemble subitizing and Object subitizing blocks. Colors of ensembles and objects were randomly chosen throughout the course of the experiment in order to disallow preference for a particular color.

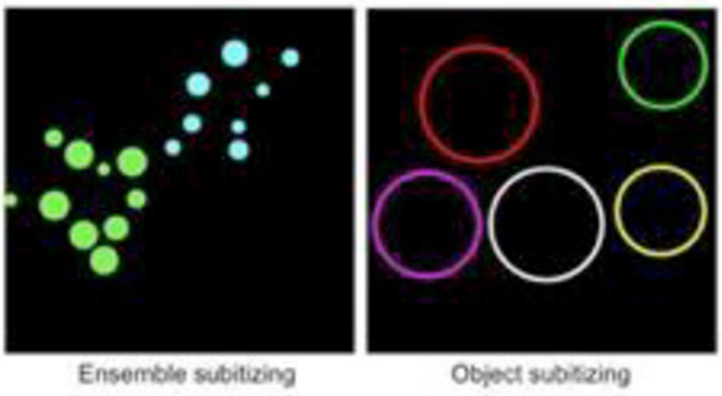


Figure 4. 1 Sample displays for Ensemble subitizing and Object subitizing tasks

Procedure

Figure 4.2 shows the sequence of a trial. After a screen indicating that the subject will begin the trial, the stimulus was displayed and remained on the screen until the subject pressed the space bar. Subjects were asked to press the space bar immediately after determining the exact number of colors in the stimulus. A mask was flashed for 200 msec after they pressed the space bar, followed by a dialog box prompting the subject to enter the numerosity. Accuracy and response time were recorded. The subjects completed 150 trials of Ensemble subitizing and 150 trials of Object subitizing.

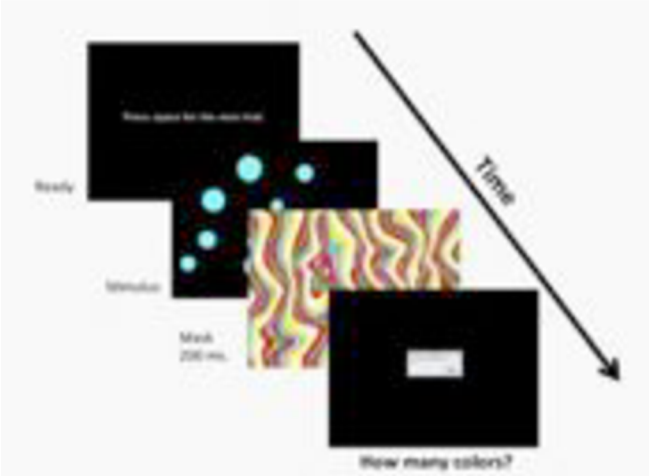


Figure 4. 2 A sample trial of Ensemble subitizing task

4.3.2 Results and Discussion

Results from subjects who had an overall accuracy of less than 90% (4 total) were discarded. Accuracy rates for the remaining subjects are displayed in Figure 4.3. Subjects were accurate with errors increasing with set size both for Ensemble subitizing and Object subitizing. For analyses of response times (RT), incorrect trials were removed from further analysis. For the correct trials, we then removed outlier RT's for each subject, for each condition according to a procedure advocated by Selsts and Jolicoeur (1994). Subitizing elbows – the position at which subitizing RTs transitioned to counting

RTs (also known as subitizing capacity) – were found using nonlinear least squares regression with the constraint that the first slope must be small and positive, the second slope must be greater than the first slope and the two lines must intersect – as has been described elsewhere (C. S. Green & Bavelier, 2006). The slopes comprising the elbow curve for Ensemble subitizing and Object subitizing were nearly identical ($F(1,25) = 0.31$, $p = 0.58$), supporting the similarity of ensemble and object subitizing. The finding that subitizing ensembles results in an almost identical subitizing elbow curve to that of objects suggests that the visual indexing procedure is identical, and visual selective attention is carried out in the same manner for both groups of similar units as for individual units themselves.

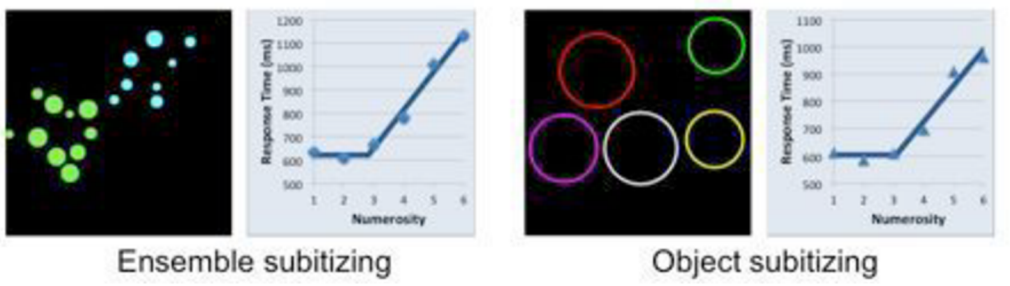


Figure 4. 3 Avarege RT as a function of the numerosity

4.4 EXPERIMENT E: SUBITIZING ENSEMBLES AND APPROXIMATE NUMBER

The results from Experiment D suggest that human observers can select ensembles as units for enumeration. However, one might argue that it is unclear whether the observers in fact represented ensembles as groups or simply sampled each group coarsely in order to report the number of colors that were presented in the visual array –

without encoding any ensemble information from each group. Therefore, Experiment E was conducted as a simple follow-up to ensure that the observers represented ensembles rather than only picking up color information from a few of the individuals or a coarse coding. In Experiment E, I asked subjects to report either the number of groups (Ensemble subitizing) or the number of individual dots within an ensemble (Approximate number).

4.4.1 Experimental method

Subjects

6 naive subjects participated in the experiment for course credits. All of the subjects had normal or corrected-to-normal vision.

Apparatus and stimuli

All the aspects of stimuli were identical to those in Experiment D. In this experiment, I only used the stimuli from the Ensemble subitizing trials in which 1-6 clusters of dots with different colors were presented.

Procedure

Figure 4.4 shows the sequence of a trial. After a screen indicating that the subject will begin the trial, the stimulus was displayed. The stimulus duration was varied from 250 msec to 1000 msec. After the stimulus array, a mask was presented for 200 msec and then the subjects were prompted to report either on the number of clusters (Ensemble subitizing) or the number of elements of one of the presented clusters (Approximate number). For the Approximate number question, a probe with the color of one of the clusters appeared on the location of the cluster that the subject had to report about. The subjects were asked to enter the number of dots in that cluster by using the number pad

on the keyboard. Accuracy and response time were recorded. The subjects received 300 trials and on a half of the 300 trials they enumerated clusters (Ensemble subitizing) and on the other half they estimated the number of elements in one cluster (Approximate number).

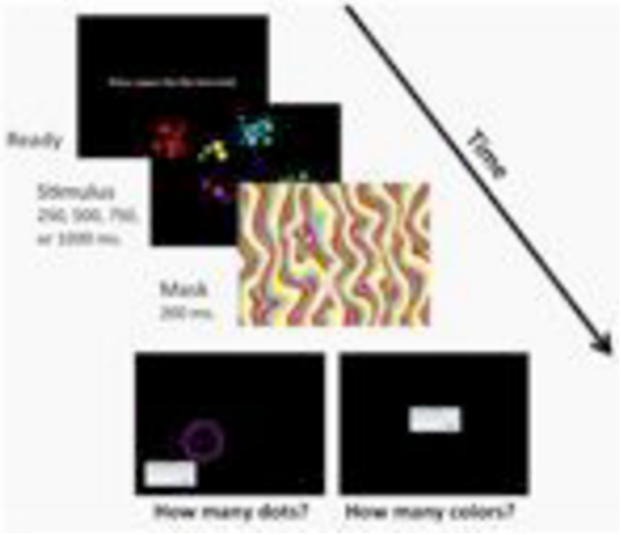


Figure 4. 4 A sample of Experiment E

4.4.2 Results and Discussion

Since I varied the display time in Experiment E, I focus my analysis on accuracy measures rather than RT. Figure 4.5a shows the proportion error of the subjects' responses on the ensemble subitizing task. Subitizing performance plotted as a function of display time reveals the time course of subitization of ensembles, with robust elbows at 500 and 750 msec. At 250 msec, with the mask I used here, subjects were not especially accurate; and at 1000 msec it appears that subjects began to focus on enumeration with a reduction in error for displays with 5 and 6 clusters. At all display times, the proportion error increased with the increasing number of the ensembles.

To address the main question of interest, I determined whether the subjects were actually able to extract ensemble feature information, i.e., the approximate number of elements in each cluster. Therefore, I examined the relationship between the subjects' accuracy on ensemble subitizing (Figure 4.5a) and accuracy on estimation of approximate number of elements within an ensemble. Figure 4.5b shows the scatter plot between the accuracy on the ensemble subitizing (X-axis) and the accuracy on the estimation of number of elements within an ensemble (Y-axis). I found a high correlation between these two different measures of accuracy for ensemble subitizing and ensemble feature representation ($r = 0.75$, $p < .01$). The result that ensemble feature precision is highly correlated with subitizing success suggests that each cluster gives rise to ensemble feature representation. In addition, the results also suggest that when subjects successfully subitized, they also successfully represented the ensemble feature (e.g., approximate number) for each cluster. Therefore, these results further support the conclusion from Experiment D that an ensemble can be a unit for selection, and when an ensemble is selected, one can extract approximate number from the selected ensemble.

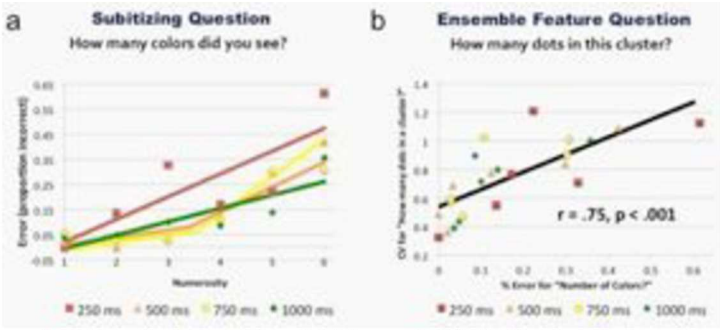


Figure 4. 5 (a) Proportion error for Ensemble subitizing task (b) CV for Approximate number task

CHAPTER 5. ENSEMBLES IN WORKING MEMORY: EACH ENSEMBLE IS CONSOLIDATED INTO VISUAL WORKING MEMORY IN AN ALL-OR-NONE FASHION

5.1 SYNOPSIS

Chapter 5 explores how ensemble representations are encoded and stored in memory. Specifically, I sought to determine if ensemble representations flexibly change in precision such that representations become more precise with more encoding time and with lower memory load. To address this question, I varied both stimulus duration (e.g. 33, 66, 99, 132, and 198 msec) and the number of ensembles presented in the visual array (e.g., 1, 3, or 6 sets). Here I introduce a new approach to measuring the precision of internal representation in a more robust manner, by empirically measuring an individual subject's guessing pattern and by using this estimate of the subject's guessing pattern to aid in filtering out guess trials from the mixture of guess and regular responses on the trials of interest (i.e., trials where a stimulus was actually presented). The results of this modeling approach suggest that, when guess trials are appropriately filtered out, the precision of internal representations does not appear to improve over time or with lower memory load. This supports the construct of a fixed precision of visual working memory representations for the ensemble feature approximate number.

5.2 BACKGROUND

5.2.1 Fixed or flexible precision of visual working memory representations

With every movement of our eyes or shift of attention, the visual environment is constantly changing and each snapshot of the visual world is overflowing with information about various features, objects, and collections. Given the dynamic and complex nature of the visual world, processing time (e.g., the amount of time we can dedicate to gathering evidence to represent the visual world) and capacity limitations (e.g., the total amount of information that can be processed or stored at a given time) have been widely discussed as major factors that limit the resolution of human visual representation and cognition more generally. For example, many have suggested that the precision and accuracy of our visual representations improve progressively such that they become more precise and refined when we are given longer periods of time to view a stimulus (e.g., Carrasco & McElree, 2001; Gegenfurtner & Sperling, 1993; Grill-Spector & Kanwisher, 2005; Liu & Jiang, 2005; McElree & Carrasco, 1999; Vogel, Woodman, & Luck, 2006) or a smaller number of visual items to process (e.g., Alvarez & Franconeri, 2007; Bays & Husain, 2008; Dobkins & Bosworth, 2001; Franconeri et al., 2007; Luck & Vogel, 1997; Palmer et al., 1993; Palmer, 1990; Sakai et al., 2007)

Consistent with the intuition that our visual representations can improve with increased viewing time and with reduced cognitive load, a number of authors have recently championed a 'flexible resource' model of visual working memory (VWM) where representations are allowed to have a flexible, or graded, amount of precision (sometimes called 'variable precision') (Bays, Catalao, & Husain, 2009; Bays & Husain, 2008; Palmer, 1990; van den Berg, Shin, Chou, George, & Ma, 2012; Wilken & Ma,

2004). However, empirical findings of better performance in visual tasks involving longer display times and fewer visual items could occur given a fixed resolution, discrete system of visual representations - e.g., because observers might be able to process and store more discrete bits of information from a visual scene as viewing time is increased or item-load is reduced (e.g., Anderson, Vogel, & Awh, 2011; Pashler, 1988; Province & Rouder, 2012; Rouder et al., 2008; Zhang & Luck, 2008). And, when time is severely limited or too much information needs to be processed at the same time, human observers might encode a portion of information from the scene (at a fixed precision) and then rely on strategies to fill-in information about any unprocessed regions or items (e.g., strategic guessing).

Because behavioral responses in a visual task likely reflect a mixture of responses based on internal representations and those based on strategic guesses it has become crucial to sort a subject's responses into these two classes (or a blending of these classes) in order to determine if internal representations are indeed 'flexible' (i.e., with variable precision across time and load) or 'fixed' (i.e., at a specific level of precision). To date, authors in the vision sciences, and throughout psychophysics more generally, have included parameters in their models to estimate the frequency of random guesses that are not responsive to the signal (e.g., the inclusion of 'lapse parameters' in models of signal-detection tasks). This requires assuming that human guessing is unstrategic and uniformly distributed across all possible signals. But such assumptions are likely to be untrue of real human observers, and any mismatch between these assumptions and actual human performance will obfuscate attempts to estimate the precision of internal visual representations.

Here, I directly measure the guessing patterns of individual observers in the absence of a physical stimulus during visual tasks. I find that guessing is never random nor uniformly distributed - calling into question the assumptions of many previous modeling efforts (e.g., Fougnie & Alvarez, 2011; Green & Swets, 1966; Halberda & Feigenson, 2008; Ludwig & Rhys Davies, 2011; Rouder et al., 2008; Zhang & Luck, 2008, 2011). I then use each subject's idiosyncratic guessing pattern to better identify possible guesses that may have occurred on trials where a signal was presented - i.e., I determine a likelihood for each behavioral response indicating an estimate of the probability that the response was based on the subject's internal representation of the visual stimulus as opposed to their own personal guessing strategy. This likelihood (a continuous likelihood ranging from 0 'pure guess' to 1 'pure internal representation') can also be used to estimate indeterminate blends (e.g., likelihood ~.5) where the subject might have used either their internal representation, a strategic guess, or a blending of the two. I then use the likelihood from each trial to weight each trial's contribution to a modeled estimate of the internal precision of the subject's visual representations. This allows me to determine whether visual representations have a 'flexible' or 'fixed' precision, while controlling for strategic guessing, both across viewing times and item-loads. To date, this approach has not been used, and previous authors have either assumed that guesses are uniform and unstrategic (Fougnie & Alvarez, 2011; Green & Swets, 1966; Halberda & Feigenson, 2008; Ludwid & Davies, 2011; Rouder et al., 2008; Zhang & Luck, 2008; 2011) or - in the most extreme case - that human subjects never actually guess during behavioral tasks (van den Berg et al., 2012).

In Chapter 5, I demonstrate that ensemble representations are encoded and stored at a fixed, not flexible, precision and that behavioral responses within a trial are either based on this fixed internal representation or on a strategic guess, and rarely if ever on a blending of the two. For coherence, I chose to focus on approximate number and average size throughout this dissertation; however, though not included in this dissertation, I also tested other visual features for individual objects (Color, Orientation, and Length) in separate experiments outside of this dissertation. In all of these experiments, I used this same approach to model human responses and I also found the same result, i.e., the precision of internal representations in VWM did not change over time and with different item loads. The same pattern that I observed from representing ensemble (e.g., approximate number) and individual features from single items (e.g., Color, Orientation, and Length) suggests that approximate number can be extracted in the same manner as other basic visual features from individual objects. Going beyond ensemble representation, this chapter will more generally propose that visual representations (including ensembles) are best understood as emerging from a detect-or-guess fixed resolution system - a finding that is consistent with diffusion-to-criterion models of the encoding of motion (Ditterich, 2006) and orientation information (Ludwig & Rhys Davies, 2011), and is consistent with recent results suggesting that humans rely on a recall-or-guess retrieval of representations from long-term memory (Province & Rouder, 2012). This recall-or-guess model is also consistent with recent studies on the decay of working memory representations (Zhang & Luck, 2009) showing that precision of the stored representations decay in an all-or-none manner, rather than gradually decreasing over time.

The results presented in this chapter (and the results from the work I've done on individual object features that does not appear in this dissertation) bear on recent claims that VWM relies on flexible representations - claims that were based on studies which assumed human guessing to be nonstrategic and uniformly distributed (Bays et al., 2009; Palmer, 1990). In the most extreme case, authors of previous modeling efforts have gone so far as to suggest that humans never guess at all (van den Berg et al., 2012; Wilken & Ma, 2004). The work in this chapter presents empirical evidence that these assumptions are inappropriate and that the inaccuracy of these assumptions can lead psychological models to falsely suggest that visual representations are flexible.

5.3 EXPERIMENT F: MEASURING ESTIMATION BIAS AND INTERNAL PRECISION

5.3.1 Experimental method

Subjects

10 subjects participated in the experiment. All of the subjects had normal or corrected-to-normal vision. The subjects received monetary compensation for participation.

Apparatus and stimuli

Stimuli were generated using MATLAB software together with the Psychophysics Toolbox extensions (37, 38) and displayed on a LCD monitor with a grey background driven by a Macintosh iMac computer (viewable area was a gray central square window with a 17'' diagonal). The subjects were seated approximately 50 cm from the screen, and viewed the display binocularly. At this viewing distance, each pixel

was approximately 0.04° of visual angle. In each stimulus array, 1 to 6 clusters of multiple dots with different sizes were presented, each within its own invisible circle with a radius of 1.8°. The positions of the invisible circles were randomly chosen from a set of eight locations equally spaced along the circular display region at an eccentricity of 5.3°. The number of dots in each cluster was varied from 5 to 35 and the size of each dot was randomly chosen from a range between 0.12° and 0.23°. Dots within each invisible circle were randomly located with the constraint that they never overlap. In the response array, the response number scale was a linear, evenly divided bar with a number marked from 0 to 45. The number scale was 1.2° thick and was located in the bottom of the memory array.

Procedure

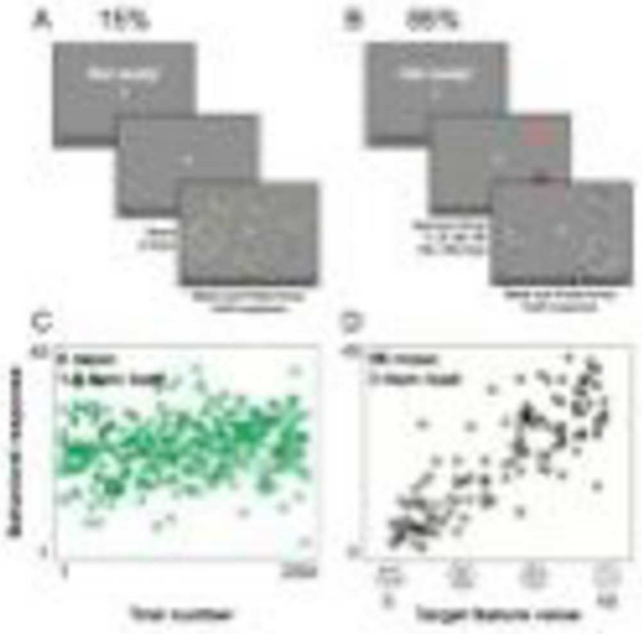


Figure 5. 1 A sample trial of Experiment F. The column A indicates 0-msec trials and the column B indicates the regular trials with longer durations

Figure 5.1 illustrates example guess and test trials. Each trial consisted of a stimulus array with a varying duration (0, 33, 66, 99, 132, or 198 msec) followed by a mask and probe display that remained present until a response was made. The stimulus array consisted of one, three, or six collections of multiple dots. Feature value of each item was randomly selected from a uniform distribution (e.g., a target value between 5 and 35 dots). The probe display contained a white circle to indicate which cluster subjects should recall and the linear response scale ranging from 0 to 45 on which they clicked the numerosity value of the probed cluster using the computer mouse. Accuracy was stressed, but subjects were asked to estimate and to make a response as soon as possible. The duration of the memory array and the array size (e.g., 1, 3, or 6 cluster) was randomly varied and subjects were informed when they could not recall the stimulus, they should make their best guess. There were eight separate blocks consisting of 198 test trials (total 1584 trials per subject) and subjects received 30 practice trials before the test trials.

5.3.2 Modeling: A new approach to mixture modeling

My mixture model consists of a model for guessing that is built upon empirically measured responses from an individual observer on trials with 0-msec-stimulus duration and a model for internal representation for approximate number.

Model structure

In the model for internal representation, two free parameters were defined and fitted via the standard maximum likelihood estimation (MLE) procedure: P_{int} reflecting the mean probability of the data points being drawn from an internal representation of the

stimulus and CV_{int} which inversely reflects the precision of internal representation of approximate number (i.e., smaller CV_{int} indicates more precise internal representations).

The first parameter of P_{int} estimates the overall proportion of an observer's responses that are likely to have been drawn from an internal representation of the presented stimulus image, as opposed to guessing. Thus the compliment of P_{int} would reflect the guess rate – that is, the degree to which an observer relied on their non-visual guess strategy rather than consulting their internal representation.

The second parameter, CV_{int} , is a parameter that determines the normalized standard deviation of Gaussian distributions, describing inherent noise of the internal representations. Due to scalar variability, the standard deviation of the Gaussian distribution for each numerosity is a function of the stimulus numerosity such that SD increases linearly with the stimulus. CV_{int} reflects the overall precision of the internal representations, independent of the presented numerosity, via normalization by the mean of the actual magnitude. The model's estimate of CV_{int} was determined by weighting the contribution of each observation by the estimated likelihood value for each data point – i.e., the likelihood that the response was drawn from an internal representation of the stimulus (P_{int} for each data point). In this way, using each subject's personal guessing strategy to determine the likelihood that a response was generated based on an internal representation of the stimulus (versus being drawn from their own personal guess strategy) allows observations that were more likely drawn from an internal representation to contribute more to the MLE estimation of CV_{int}.

The model for guessing does not simply assume a uniform distribution but rather involves empirical measurement of individual subject's guessing pattern based on their

responses from trials with 0-ms-stimulus-duration. I estimated the distribution of each subject's guess responses on the trials with 0-msec-stimulus duration by using Kernel density estimation, which is a non-parametric way of estimating the probability density function of a random variable (Rosenblatt, 1956; Wasserman, 2006).

Algorithm

From each data point on an individual trial, the model evaluates the probability of the single data point of subject's response being drawn from internal representation as opposed to being drawn from guessing. The probability of a single data point (Xi, Yi), if it follows internal representation, is defined as:

$$\Pr(Y_i|X_i, CV_{int}) = \frac{1}{\sqrt{2\pi X_i \cdot CV_{int}{}^2}} \exp\left\{-\frac{1}{2CV_{int}{}^2}\left[Y_i - X_i{}^{0.9}\right]^2\right\}, \quad [7]$$

where X_i is the actual numerosity of dots presented in a stimulus array, Y_i is a subject's reported value, , and CV_{int} is the coefficient of variation (CV) of the internal representation. The mapping function between the actual magnitude and the reported value was defined as a power function, $Y = X^{\beta}$ and here the exponent β was fixed to be 0.9 based on the previous findings that the representation of numerosity follows a power function with an exponent typically around 0.85-0.95 (e.g., Krueger, 1984). CV_{int} is the normalized standard deviation in order to account for the known scalar variability in representations of approximate number.

The probability of a single data point (Xi, Yi), if it follows a distribution of guess responses, is defined from the probability density function calculated by kernel density estimation. For a subjects' guess responses (X_1,X_2, ..., X_i) collected from trials with 0-msec-stimulus duration , the kernel density estimator is defined to be

$$f_n(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{n} K\left(\frac{x - x_i}{h}\right), \qquad [8]$$

where n is the number of data points and h is a smoothing parameter called the bandwidth and K is the kernel function defined as the standard normal distribution for internal representation of numerosity. As with Kernal density estimation, the choice of kernel K is not crucial, but the choice of bandwidth h is important (Wasserman, 2005); therefore h was chosen by the cross-validation method for each data set (Rudemo, 1982).

For the maximum likelihood estimation (MLE), the model had a latent variable Z_i in which $Z_i = 1$, if a given data point Y_i is drawn from internal representation and $Z_i = 0$, if a given data point Y_i is drawn from pure guessing. The latent variable Z_i is dependent on a hyper parameter τ which denotes the overall probability of data points being drawn from internal representation such that:

$$\Pr(Z_i|\tau) = \tau^{Z_i}(1-\tau)^{1-Z_i}, \qquad [9]$$

where a prior for τ was defined as a beta distribution.

Using the iterative MLE procedure, the model provides values for CV_{int} and the latent variable Z (a vector of $Z_1 \ldots Z_n$) that maximizes the expected log-likelihood. As the final outcome from this procedure, the converged latent variable Z_i becomes continuous ranging from 0 (i.e., definitely drawn from guessing) to 1 (i.e., definitely drawn from internal representation). It is worth noting that Z_i values were not binary (i.e., either 0 or 1) but continuous between 0 and 1 after multiple iterations, providing a continuous probability value that indicates the degree to which a given data point has been drawn from internal representation as opposed to being drawn from guessing. Finally, P_{int} was

determined by averaging the latent variable Z to indicate the overall proportion of regular responses from internal representation within a set of responses at each stimulus duration.

5.3.3 Results and Discussion

Human guess responses are not random nor uniformly distributed

Authors of previous studies have assumed that human guesses are unstrategic, random, and uniformly distributed across all possible target feature values, but have not directly measured guessing (Fougnie & Alvarez, 2011; Green & Swets, 1966; Halberda & Feigenson, 2008; Ludwid & Davies, 2011; Rouder et al., 2008; Zhang & Luck, 2008; 2011). Here, I measured guessing empirically and we find that these assumptions are inappropriate. Figure 5.2 presents the 0 msec responses from one representative subject in along with the corresponding histogram and probability density functions derived from these guessing trials. The histogram and PDF in Figure 5.2 were formed by collapsing across all 0 msec trials for one subject. The high regions indicate regions of frequent guesses while the lows indicate regions of rare guesses. All subjects tended to avoide guessing at the highest and lowest values in the response range and instead focused the majority of their guesses in the middle of the range. Importantly, this guessing strategy did not reflect the trials subjects saw during the task, as trials were selected from a uniform distribution. Rather, the unimodal guessing strategy of subject appears to depend on their attempt to avoid large errors in estimation – i.e., a strategic decision.

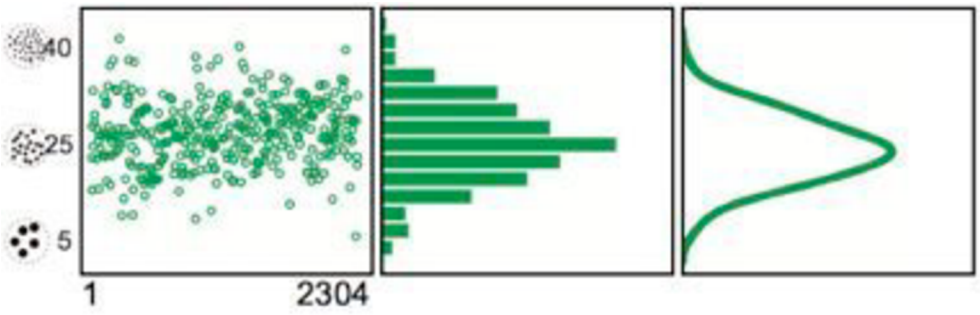


Figure 5. 2 One subject's responses collected only from the 0-msec trials

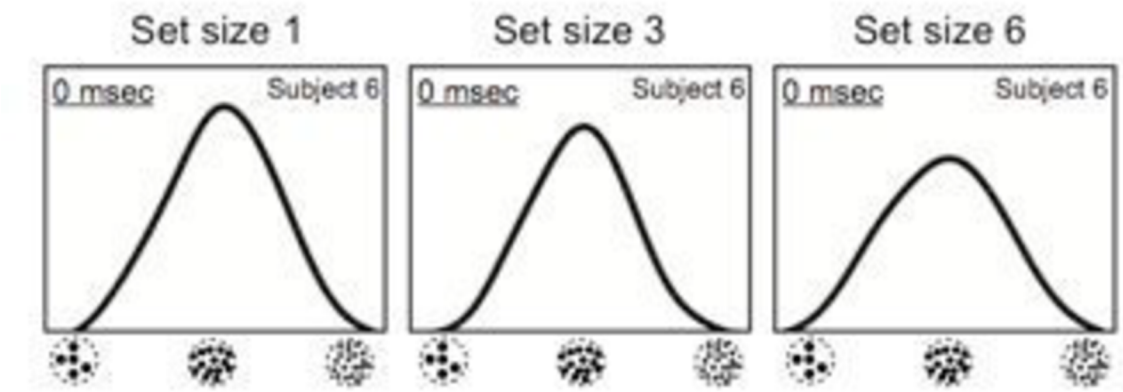


Figure 5. 3 Guessing pattern across set size (example from one subject)

I found that guessing patterns were consistent across trial number and across set sizes (i.e.,1, 3, or 6 masks appearing after the 0 msec stimulus display; Figure 5.3), and I collapsed all guessing trials into a single estimate of the guessing pattern for each subject. Thus, the highs and lows in the histograms and probability density functions result from collapsing across all 0 msec trials according to feature value; with the highs corresponding to feature values that the observer favored when guessing (i.e., frequent guess response regions) and the lows corresponding to feature values that were dis-preferred (e.g., rare guess response regions). The probability density function in Figure

5.2 was generated by nonparametric kernel density estimation (Rosenblatt, 1956; Wasserman, 2005). A non-parametric Kolmogorov-Smirnov test was performed to compare the guessing pattern from each subject with a null, uniform distribution and it confirmed non-uniformity of subjects' guessing distributions for all subjects (all p's < .01). These results suggest that human guessing is neither random nor uniformly distributed in approximate number tasks (NB, I also found non-uniformity for guessing in every visual feature I investigated, e.g., Color, Size, Orientation). The peaks of both the histogram and the probability density function at the center of the range (e.g., around 20-25) suggest that the observers favored the intermediate values when guessing but they avoided making guesses near the endpoints of the magnitude scale (e.g., near 0 or near 45) - in an apparent attempt to reduce error variability. This pattern was highly consistent across subjects. It seems that the subjects relied on a similar strategy of making guess responses toward the center of the linear number scale in order to minimize overall error. This suggests that the subjects make guess responses in an educated, strategic way to optimize their outcome instead of making non-informative random responses. The empirically measured guessing pattern was highly consistent across individual subjects. This finding suggests that subjects maintained shared expectations and knowledge about the task and visual feature (Numerosity) and brought these to bear when making strategic non-uniform guesses. The actual numerosity of the dots for each cluster was determined from uniform distribution of stimulus value throughout the task, suggesting that the non-uniform guessing patterns observed on 0 msec trials are not simply a reflection of the particular trials that subjects saw, but rather reflect subjects' own strategic decisions and knowledge of the task (e.g., an attempt to reduce error variability by focusing guesses

towards the center of magnitude scales). These results suggest that any attempt to estimate the bias and the precision of internal representations must include an empirical estimate of non-uniform guessing in order to effectively remove guesses from the dataset. Given that numerosity representation is just one example of a feature dimension with scalar variability – i.e., the same basic structure shared by the vast majority of psychological representations (e.g., brightness, loudness, felt electric shock, odor concentration, finger spacing etc., Teghtsoonian, 1971) – the present results draw attention to the possibility that the quite ubiquitous practice of including non-strategic 'lapse' parameters in psychophysical models that assume uniform, unstrategic guessing may not be accurately capturing human performance.

Representations of ensembles are encoded and stored in memory in an all-or-none manner

For the fidelity of internal representations, it is necessary to consider CV_{int}, the normalized version of standard deviation of Gaussian distributions fitted to human responses. Using the empirically-determined guessing pattern for each subject (e.g., Figure 5.2) and an appropriate model of internal precision (Figure 5.4), I ran an iterative mixture modeling procedure using maximum likelihood estimation (MLE) on each subject's raw response data to recover three components - the probability of answering based on the internal representation (P_{int}) and the observed precision of visually-guided responses (CV_{int}). This iterative MLE procedure estimated the likelihood for each trial that the subject's response was based on an internal representation of the target stimulus

(versus being drawn from the probability density function for their personal guessing pattern, or a blending of the two).

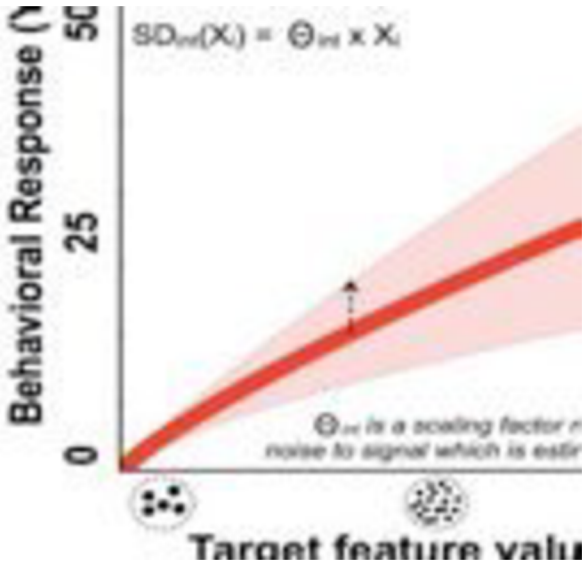


Figure 5. 4 Model for internal representation of approximate number

The output of this approach is a likelihood for each trial. The average value across these likelihoods is an estimate of the probability that the subject will respond based on an internal representation of the target (i.e., P_{int}). And, using the likelihoods as weights, every trial contributes to an estimate of the free parameter that specifies the internal precision (CV_{int}) of the visual representations. In Figure 5.5, the simulation results demonstrate that this mixture modeling approach can recover the original parameters used to create a wide variety of datasets.

Example outcomes for one subject appear in Figure 5.6. In Figure 5.6, each circle is a single trial color-coded according to the model's estimated likelihood that the subject produced this response based on an internal representation of the target (red), a guess (green), or a blending of the two (brownish). The plots in Figure 5.6 represent all of the behavioral trials from one representative subject for each item load and display duration

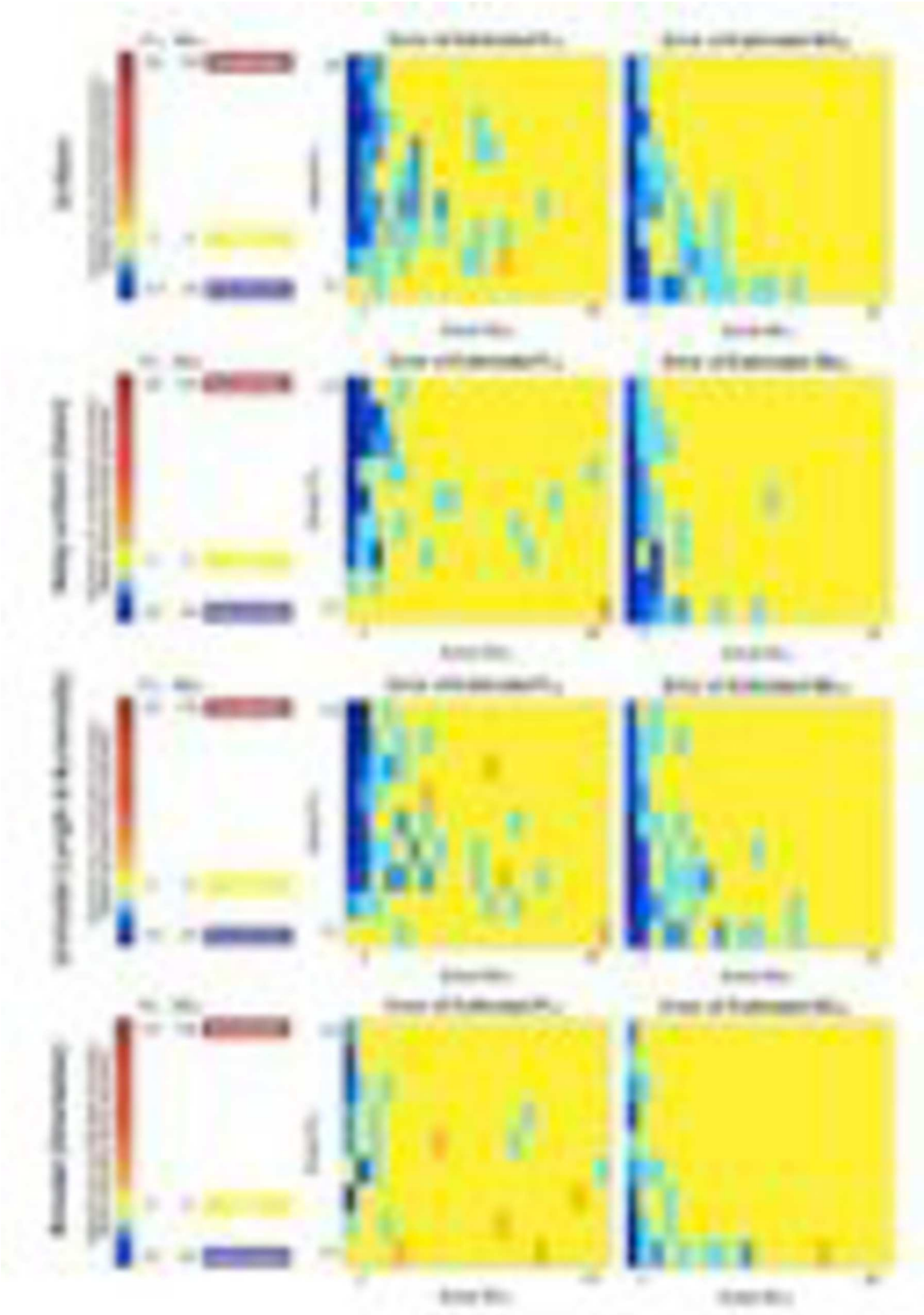


Figure 5. 5 Simulation results

and there are roughly an equal number of trials in each plot - this allows the reader to investigate the ratio of red-ish to green-ish circles in order to understand how representations and behavior may change over time (e.g., 33 to 198 msc) and over item load (e.g., 1, 3, and 6 sets).

The iterative MLE procedure was run separately for each subject and for each condition of interest (i.e., 33, 66, 99, 132, 198 msec; 1, 3, 6 items). This means that a separate MLE value for e.g., P_{int} and CV_{int} is generated for each of the plots in Figure 5.6 allowing one to ask whether the probability that a subject answers based on an internal representation increases over display time (e.g., increasing red and decreasing green circles in the plots as display time increases from 33 to 198 msec) and whether the fidelity of internal representations increase either over increasing display times or reduced item loads (e.g., reducing spread in the shape of the band of the red circles across time or load).

The probability of responding based on an internal representation, P_{int}, increased over display times as subjects tended to rely more and more on internal representations as they were given more time to view the stimulus (e.g., green circles reducing across display times in Figure 5.6). This can also be seen in Figure 5.7, which presents histograms for the likelihood values displayed in Figure 5.6. In Figure 5.7 one can see the shift in likelihood values as the majority of responses appear to be drawn from the subject's guess distribution (i.e., green bars) at shorter display times while the majority of responses appear to be drawn from the subject's internal representation of the stimulus at longer display times (i.e., red bars).

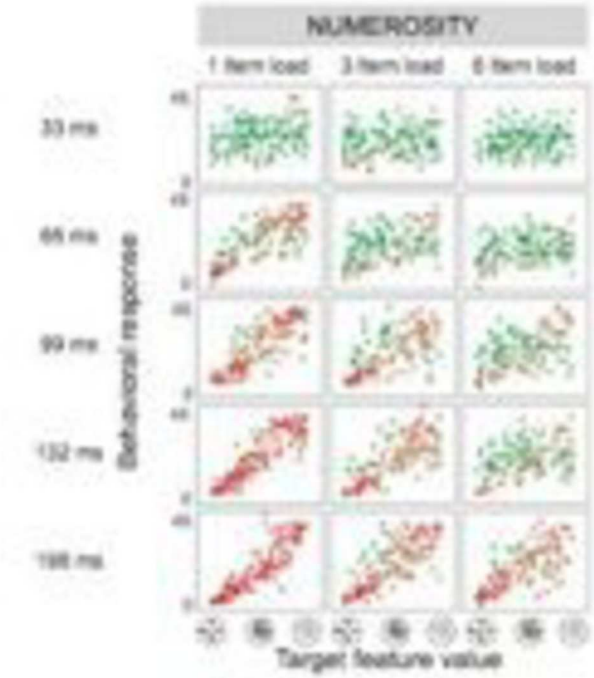


Figure 5. 6 Model results: each dot indicates likelihood of being drown from internal representation (RED) as opposed to from guesing (GREEN)

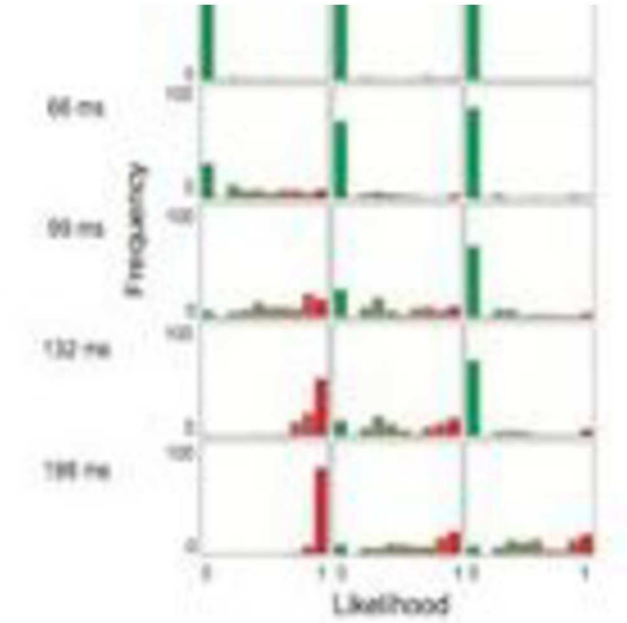


Figure 5. 7 Model results: histogram of the likelihood values from Figure 5.6

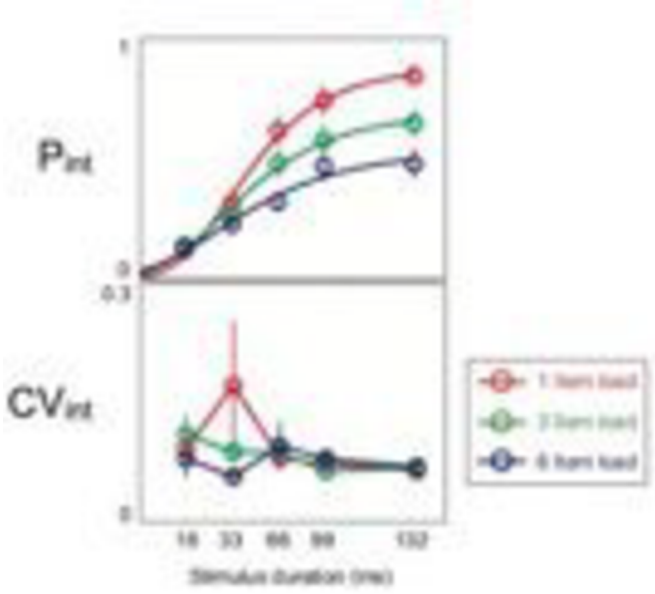


Figure 5. 8 Summary results, averaged across subjects

Using display times to estimate changes in representations over time, Figure 5.8 presents subject means ±SE for P_{int}, and CV_{int}, across display times separated by item load (i.e., 1, 3, 6 items). This is a digestion of how the two fitted parameters (P_{int} and CV_{int}) change across viewing time and item load. The increase in P_{int} across display times was statistically significant ($p < .01$). The growth curves fitted for each feature and item load suggest a rapid increase in participants tendency to rely on their internal representations. The growth curve was fitted to the estimated P_{int} for each item load by the logistic curve as:

$$f(t) = a \cdot \exp(-b \cdot \exp(-ct)) \qquad [10],$$

where t is the time point, a is the asymptotic level of performance, and b and c are shape parameters for the curve (Gompertz, 1825). From the growth curve fitted to P_{int}, I obtained both the time point at which P_{int} approached the upper asymptote and the time point at which P_{int} started increasing sharply (i.e., an elbow point). Figure 5.9 summarizes

the asymptotic level of P_{int}, the time point at which P_{int} reached the asymptotic level, and the elbow point in which P_{int} started increasing sharply. These parameters from the growth curve fits suggest that the improvement in P_{int} over time was completed within a very short time window such that P_{int} reached the asymptote very fast (e.g., less than 140 msec for all item load). In addition, the elbow point of the growth curve in which P_{int} started increasing sharply reflects the time point when observers started consulting their internal representation rather than making non-visual guess responses. Therefore, the elbow point would reflect the time point at which the internal evidence about the numerosity of the dots in the visual arrays has been accumulated enough to drive decisions based on the internal representation. Elbow points located earlier than 100 msec suggest that visual information can be accumulated and used to drive decisions very rapidly, which is consistent with other recent finding of detect-or-guess models (Ludwig & Davies, 2011).

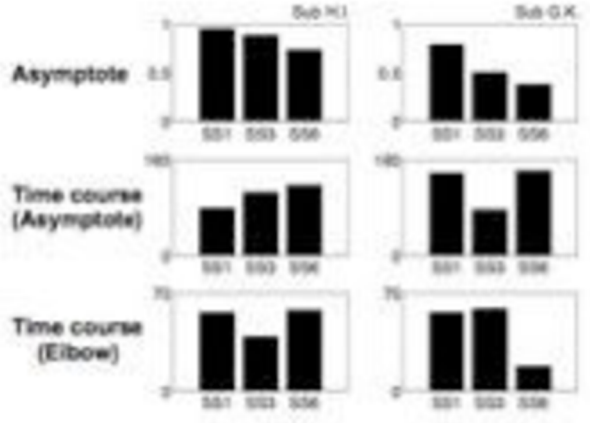


Figure 5. 9 Fitted parameters for the growth curve (examples of two subjects)

I next sought to address whether internal representations of approximate number of sets are 'flexible' or 'fixed' in precision. This can be evaluated by considering the

parameter values for CV_{int}, which index the fidelity of the internal visual representations. If visual representations have a flexible precision that can improve across display times (e.g., as the viewer gathers more visual evidence) or across decreasing item load (e.g., as the viewer has fewer items they need to process) then we should observe a significant increase in CV_{int} as item load increases (e.g., 6-item trials would result in significantly higher CV_{int} (i.e., worse precision and more underestimation) than 1-item trials) and as display time decreases (e.g., 33 msec displays would result in significantly higher CV_{int} than 198 msec displays). This is not the pattern I observe (Figure 5.8). Rather, once guessing trials are effectively removed and weighted by the likelihood function, I find little to no evidence for changes in CV_{int} across time and item load.

These results are consistent with some other recent results. In some cases, proponents of drift-diffusion models of cognition (DDM) have suggested that observers do not benefit from prolonged accumulation of perceptual evidence - beyond the point when the accumulation process reaches the threshold for a perceptual decision (Ratcliff, Van Zandt, & McKoon, 1999; Ratcliff, 1978). Our observed increases in P_{int} with increasing display times are consistent with diffusion towards a bound and with DDM in general. In addition, other work has suggested that subjects may be ignorant of perceptual signals during the perceptual accumulation process - until this process crosses the detection/decision threshold (Ratcliff, 1978; Ratcliff, Van Zandt, & McKoon, 1999). This is consistent with the current findings of no change in CV_{int} with decreasing display times. That is, even at the briefest display times, where very few trials appear to involve responses based on an internal representation, the precision of the internal representations appear to be fixed and constant. This result is also consistent with recent findings

suggesting that the precision of internal representations might be highly limited in their ability to evolve and be refined over time (Ludwig & Davies, 2011).

There are other recent suggestions that VWM representations may have a ‘fixed‘ rather than ‘flexible‘ precision, but that precision may appear higher at lower item loads do to a blending of information from multiple samples (e.g., “slots plus averaging” Zhang & Luck, 2008; 2011; also see van den Berg et al., 2012). While our results support the notion of fixed precision, our empirical work suggests that there is also no change in the precision of internal representations over time and across set sizes.

Chapter 5 demonstrates that the precision of internal representation of approximate number is fixed, not flexibly changing over time and with different set sizes. These results suggest that ensemble representations are encoded and stored in memory in a discrete and all-or-none manner.

Outside of the work presented in this dissertation, I also tested different visual features using the same experimental paradigm and the same modeling approach. These results returned converging evidence that the precision of visual internal representations does not change but remains fixed when there is more time to process and there is less information load. This converging evidence further confirms that ensemble representations such as approximate number behave much like any basic visual features such as color, orientation, or line length. In addition, as a more general conclusion, results from this broader line of research support a general framework of human visual representations: internal visual representations are fixed in precision and encoded and stored in an all-or-none manner.

It would seem to be commonsensical that the longer and 'harder' you look at something, the better you are able to represent it. Likewise, one might expect that the fewer items we process or actively remember at one time, the more accurately we will represent them. However, rather than resting on a foundation of 'flexible' precision or changing cognitive 'resources', such experiences might arise from an emerging trade-off between time, items, and the allocation of limited 'fixed' precision representations - along with the strategic benefits of nonrandom, non-uniform guessing. My work suggests that, when we are asked to make judgments about visually presented items and collections, our impression of gradually evolving/improving visual representations with variable levels of precision may be a grand illusion.

CHAPTER 6. GENERAL DISCUSSION

My dissertation research explores ensemble visual representations through perception, attention, and memory. Chapter 2 presented evidence that extraction of the ensemble feature average size is unlikely to require segmentation or sampling of individual items. The results of Experiment A suggest that representing ensembles (e.g. mean size) is more likely to rely on a distinct mechanism from that involved in representing individual objects. The mechanisms supporting ensemble representations appear to be more efficient than the mechanisms for representing individual objects and are subject to a lower level of internal noise. This is consistent with previous findings showing that the fidelity of ensemble representations are more resilient to forgetting and more robust than those for individual items (Alvarez & Oliva, 2008; Ariely, 2001; Chong & Treisman, 2003). Unlike representations of individual objects, ensemble representations appear to be rely on fast pooling processes.

In Chapter 3, I investigated one of the possible candidate algorithms that may also support the representation of ensembles: an early grouping process that gives rise to set-based representations of groups of objects. Set-based representations require individual objects to be bound into a group, providing another type of higher-order representation. Experiment B presented a new approach to providing a quantitative description of set-based representations by spatial grouping. Based on the results from Experiment B, human observers appear to rely on principles for perceptual grouping based on proximity that are similar to those incorporated in the k-means clustering algorithm from computer vision. Specifically, the estimated maximal grouping window size for human observers was highly consistent across observers and across dot arrays, suggesting that humans rely

on a default grouping distance in the vicinity of 4-degrees of visual angle. The estimated grouping window size remained constant across various numerosities of items and across different durations of visual array presentation, suggesting that grouping sets of objects is likely a parallel process rather than a serial process. Experiment C further examined a potential connection between parallel grouping and representing ensembles and showed that the pattern of set representations by spatial grouping predicted various biases of ensemble representations. These results suggest that perceptual groups built from individuals in a fast, parallel manner can affect ensemble representations of the visual array. Specifically, the well known tendency of human observers to underestimate the number of items in a stimulus can be predicted by the extent of clustering in the array as estimated by the k-means clustering algorithm. Chapters 2 and 3 together suggest that ensemble representations are extracted in a parallel manner, distinct from representing individual objects but similar to texture perception or set-based representations by grouping.

Chapter 4 explored ensembles as units of attentional selection. I tested ensemble-based subitizing – since subitizing is known to require attentional selection to index each individual unit. Experiment D revealed that human observers can use ensembles as units for visual indexing and can subitize the number of ensemble groups in a display rather than just individual objects that make up a cluster. Additionally, Experiment E further demonstrates that the success of ensemble subitizing and the success of extraction of an ensemble feature (approximate number of elements within an ensemble) highly correlated with each other, suggesting that selection of an ensemble may be required for the

extraction of approximate number from the ensemble. Experiments D and E, therefore, together suggest that ensembles can be higher-order units for attentional selection.

The research in Chapter 5 demonstrates that ensemble representations are encoded and stored in memory in an all-or-none manner. Experiment F shows that the precision of internal representations of an ensemble feature (approximate number) does not change over time or with varying set sizes. This supports a proposal for fixed-precision of visual representations. In work outside the scope of this dissertation, I have extended this approach to other basic visual features for individual objects (Color, Orientation, and Line length). The same basic pattern was observed, suggesting that the precision of internal visual representation does not change over viewing time and item load. From Chapter 5, and from the other related experiments that used the same approach but different types of visual features, I draw a more general framework of human visual representation: the fidelity of internal visual representations is fixed and inflexible, and just like representations of visual objects, ensemble representations are also extracted, encoded, and stored in memory in a discrete and all-or-none manner.

The work in this dissertation represents an attempt to build a unified understanding of visual processing and the perceptual and cognitive mechanisms involved in ensemble feature representation. The picture that emerges from my work is one in which whole groups can function as units for perception and cognition, ensemble features of these groups are extracted rapidly and in parallel (i.e., without sub-sampling), and features are stored in memory in an all-or-none fashion at a fixed resolution. The shape of my proposal owes a debt to Gestalt psychology, and ideas of grouping and hierarchical representation that have played a central role throughout information

processing approaches to modeling the human mind. While the work in this dissertation represents, perhaps, a case study – focused on approximate number – I believe that it presents an honest and robust attempt to investigate the functioning of ensemble features from the earliest stages of visual processing (e.g., perceptual grouping) through cognition and working memory. I feel that this approach is richer than focusing on any one of these stages in isolation, and I look forward to future detailed work along similar lines for other visual features (e.g., Color, Orientation, Length).

CHAPTER 7. REFERENCES

Alvarez, G. A. (2011). Representing multiple objects as an ensemble enhances visual cognition. *Trends in Cognitive Sciences*, *15*(3), 122–131.

Alvarez, G. A., & Franconeri, S. L. (2007). How many objects can you track?: Evidence for a resource-limited attentive tracking mechanism. *Journal of Vision*, *7*(13), 1–10.

Alvarez, G. A., & Oliva, A. (2008). The representation of simple ensemble visual features outside the focus of attention. *Psychological Science*, *19*(4), 392–398.

Alvarez, G. A., & Oliva, A. (2009). Spatial ensemble statistics are efficient codes that can be represented with reduced attention. *Proceedings of the National Academy of Sciences*, *106*(18), 7345–7350.

Anderson, D. E., Vogel, E. K., & Awh, E. (2011). Precision in visual working memory reaches a stable plateau when individual item limits are exceeded. *Journal of Neuroscience*, *31*(3), 1128–1138.

Ariely, D. (2001). Seeing Sets: Representation by Statistical Properties. *Psychological Science*, *12*(2), 157–162. doi:10.1111/1467-9280.00327

Ariely, D. (2008). Better than average? When can we say that subsampling of items is better than statistical summary representations? *Perception & psychophysics*, *70*(7), 1325–1326.

Atkinson, J., Francis, M. R., & Campbell, F. W. (1976). The dependence of the visual numerosity limit on orientation, colour, and grouping in the stimulus. *Perception*, *5*(3), 335–342.

Bae, G. Y., & Flombaum, J. I. (2012). Close encounters of the distracting kind: identifying the cause of visual tracking errors. *Attention, Perception, & Psychophysics*, *74*(4), 703–715.

Baker, C. L., & Mareschal, I. (2001). Processing of second-order stimuli in the visual cortex. *Progress in brain research*, *134*, 171–191.

Balas, B., Nakano, L., & Rosenholtz, R. (2009). A summary-statistic representation in peripheral vision explains visual crowding. *Journal of Vision* , *9* (12). doi:10.1167/9.12.13

Bauer, B. (2009). Does Stevens's power law for brightness extend to perceptual brightness averaging? *The Psychological Record*, *59*, 171–186.

Bays, P. M., Catalao, R. F. G., & Husain, M. (2009). The precision of visual working memory is set by allocation of a shared resource. *Journal of Vision* , *9* (10). doi:10.1167/9.10.7

Bays, P. M., & Husain, M. (2008). Dynamic shifts of limited working memory resources in human vision. *Science*, *321*(5890), 851–854.

Beaudot, W. H. A., & Mullen, K. T. (2005). Orientation selectivity in luminance and color vision assessed using 2-d band-pass filtered spatial noise. *Vision Research*, *45*(6), 687–696.

Beck, D. M., & Kastner, S. (2009). Top-down and bottom-up mechanisms in biasing competition in the human brain. *Vision Research*, *49*(10), 1154–1165.

Bergen, J. R., & Adelson, E. H. (1988). Early vision and texture perception. *Nature*, *333*(6171), 363–364.

Bergen, J. R., & Julesz, B. (1983). Parallel versus serial processing in rapid pattern discrimination. *Nature*, *303*(5919), 696–698.

Bevan, W., Maier, R. A., & Helson, H. (1963). The Influence of Context upon the Estimation of Number. *The American journal of psychology*, *76*(3), 464–469.

Blake, A., & Marinos, C. (1990). Shape from Texture: Estimation, isotropy and moments.pdf. *Artificial Intelligence*, *90*, 323–380.

Brady, N., & Field, D. J. (2000). Local contrast in natural images: normalisation and coding efficiency. *Perception*, *29*(9), 1041–1055.

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial vision*, *10*(4), 433–436.

Broadbent, D. E. (1958). *Perception and communication*. London: Pergamon Press.

Bundesen, C., & Larsen, A. (1975). Visual transformation of size. *Journal of Experimental Psychology: Human Perception and Performance*, *1*(3), 214–220.

Burr, D., & Ross, J. (2008). A visual sense of number. *Current biology : CB*, *18*(6), 425–428.

Carrasco, M, & McElree, B. (2001). Covert attention accelerates the rate of visual information processing. *Proceedings of the National Academy of Sciences*, *98*(9), 5363–5367.

Carrasco, Marisa. (2011). Visual attention: The past 25 years. *Vision Research*, *51*(13), 1484–1525.

Cavanagh, P. (2001). Seeing the forest but not the trees. *Nature Neuroscience*, *4*(7), 673–674.

Cave, K. R., & Kosslyn, S. M. (1989). Varieties of size-specific visual selection. *Journal of Experimental Psychology: General*, *118*(2), 148–164.

Chase, W. G., & Ericsson, K. A. (1981). *Cognitive Skills and Their Acquisition - Google Books*. *Cognitive skills and their acquisition*.

Chong, S. C., & Treisman, A. (2003). Representation of statistical properties. *Vision Research*, *43*(4), 393–404.

Compton, B. J., & Logan, G. D. (1993). Evaluating a computational model of perceptual grouping by proximity. *Perception & Psychophysics*, *53*(4), 403–421. doi:10.3758/BF03206783

Corbett, J. E., Wurnitsch, N., Schwartz, A., & Whitney, D. (2012). An aftereffect of adaptation to mean size. *Visual Cognition*, *20*(2), 211–231.

Cowan, N. (2001). The magical number 4 in short-term memory: a reconsideration of mental storage capacity. *The Behavioral and brain sciences*, *24*(1), 85–87.

Cowan, Nelson, Chen, Z., & Rouder, J. N. (2004). Constant capacity in an immediate serial-recall task: a logical sequel to Miller (1956). *Psychological science*, *15*(9), 634–40. doi:10.1111/j.0956-7976.2004.00732.x

Dakin, S C, & Watt, R. J. (1997). The computation of orientation statistics from visual texture. *Vision Research*, *37*(22), 3181–3192.

Dakin, S. C. (1999). Orientation variance as a quantifier of structure in texture. *Spatial vision*, *12*(1), 1–30.

Dakin, S. C. (2001). Information limit on the spatial integration of local orientation signals. *Journal of the Optical Society of America. A, Optics, image science, and vision*, *18*(5), 1016–1026.

Dakin, Steven C, Bex, P. J., Cass, J. R., & Watt, R. J. (2009). Dissociable effects of attention and crowding on orientation averaging. *Journal of Vision*, *9*(11), 28.1–16.

De Fockert, J., & Wolfenstein, C. (2009). Rapid extraction of mean identity from sets of faces. *Quarterly journal of experimental psychology (2006)*, *62*(9), 1716–1722.

Demanins, R., Hess, R. F., Williams, C. B., & Keeble, D. R. (1999). The orientation discrimination deficit in strabismic amblyopia depends upon stimulus bandwidth. *Vision Research*, *39*(24), 4018–4031.

Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual review of neuroscience*, *18*, 193–222.

Ditterich, J. (2006). Stochastic models of decisions about motion direction: Behavior and physiology. *Neural Networks*, *19*, 981–1012.

Dobkins, K. R., & Bosworth, R. G. (2001). Effects of set-size and selective spatial attention on motion processing. *Vision Research*, *41*(12), 1501–1517.

Egeth, H. E., Leonard, C. J., & Palomares, M. (2008). The role of attention in subitizing: Is the magical number 1? *Visual Cognition*, *16*, 463–473. doi:10.1080/13506280801937939

Egly, R., Driver, J., & Rafal, R. D. (1994). Shifting visual attention between objects and locations: evidence from normal and parietal lesion subjects. *Journal of Experimental Psychology: General*, *123*(2), 161–177.

Eisinger, R., Im, H. Y., Pailian, H., & Halberda, J. (2013). Ensemble-based Change Detection. *Journal of Vision*, *13*(9), 800–800. doi:10.1167/13.9.800

Emmanouil, T. A., & Treisman, A. (2008). Dividing attention across feature dimensions in statistical processing of perceptual groups. *Perception & psychophysics*, *70*(6), 946–954.

Feigenson, L. (2008). Parallel non-verbal enumeration is constrained by a set-based limit. *Cognition*, *107*(1), 1–18. doi:10.1016/j.cognition.2007.07.006

Fougnie, D., & Alvarez, G. A. (2011). Object features fail independently in visual working memory: evidence for a probabilistic feature-store model. *Journal of Vision*, *11*(12).

Franconeri, S. L., Alvarez, G. A., & Enns, J. T. (2007). How many locations can be selected at once? *Journal of Experimental Psychology: Human Perception and Performance*, *33*(5), 1003–1012.

Freeman, J., & Simoncelli, E. P. (2011). Metamers of the ventral stream. *Nature Publishing Group*, *14*(9), 1195–1201.

Frith, C., & Frith, U. (1972). The solitaire illusion: An illusion of numerosity.pdf. *Perception & psychophysics*, *11*(6), 409–410.

Gegenfurtner, K. R., & Sperling, G. (1993). Information transfer in iconic memory experiments. *Journal of Experimental Psychology: Human Perception and Performance*, *19*(4), 845–866.

Ginsburg, N. (1976). Effect of item arrangement on percerved numerosity: Randomness vs regularity. *Perceptual & Motor Skills*, *43*, 663–668.

Ginsburg, N., & Goldstein, S. R. (1987). Measurement of visual cluster. *The American journal of psychology*, *100*(2), 193–203.

Gompertz, B. (1825). On the Nature of the Function Expressive of the Law of Human Mortality, and on a New Mode of Determining the Value of Life Contingencies. *Philosophical Transactions of the Royal Society of London*, *115*, 513–583. doi:10.2307/107756

Green, C. S., & Bavelier, D. (2006). Enumeration versus multiple object tracking: the case of action video game players. *Cognition*, *101*(1), 217–245.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.

Grill-Spector, K., & Kanwisher, N. (2005). Visual recognition: as soon as you know it is there, you know what it is. *Psychological Science*, *16*(2), 152–160.

Grossberg, S., & Mingolla, E. (1985). Neural dynamics of perceptual grouping: textures, boundaries, and emergent segmentations. *Perception & psychophysics*, *38*(2), 141–171.

Haberman, J., & Whitney, D. (2007a). Rapid extraction of mean emotion and gender from sets of faces. *Current biology : CB*, *17*(17), R751–3. doi:10.1016/j.cub.2007.06.039

Haberman, J., & Whitney, D. (2007b). Rapid extraction of mean emotion and gender from sets of faces. *Current Biology*, *17*(17), R751–3.

Haberman, J., & Whitney, D. (2010). The visual system discounts emotional deviants when extracting average expression. *Attention, Perception, & Psychophysics*, *72*(7), 1825–1838.

Halberda, J., & Feigenson, L. (2008). Developmental change in the acuity of the "number sense": The approximate number system in 3-, 4-, 5-, and 6-year-olds and adults. *Developmental Psychology*, *44*(5), 1457–1465.

Halberda, J., Sires, S. F., & Feigenson, L. (2006). Multiple spatially overlapping sets can be enumerated in parallel. *Psychological Science*, *17*(7), 572–576.

Heeley, D. W., Buchanan-Smith, H. M., Cromwell, J. A., & Wright, J. S. (1997). The oblique effect in orientation acuity. *Vision Research*, *37*(2), 235–242.

Hoffman, J. E. (1980). Interaction between global and local levels of a form. *Journal of Experimental Psychology: Human Perception & Performance*, *6*, 222–234.

Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of Physiology*, *160*, 106–154.

Im, H. Y., & Chong, S. C. (2009). Computation of mean size is based on perceived size. *Attention, Perception, & Psychophysics*, *71*(2), 375–384.

Im, H. Y., & Halberda, J. (2012). The effects of sampling and internal noise on the representation of ensemble average size. *Attention, Perception, & Psychophysics*, *75*(2), 278–286.

Kahneman, D., Treisman, A., & Gibbs, B. J. (1992). The reviewing of object files: object-specific integration of information. *Cognitive Psychology*, *24*(2), 175–219.

Kaufman, E. L., Lord, M. W., Reese, T. W., & Volkmann, J. (1949). The Discrimination of Visual Number. *The American journal of psychology*, *62*(4), 498–525.

Kersten, D. (1987). Predictability and redundancy of natural images. *Journal of the Optical Society of America. A, Optics and image science*, *4*(12), 2395–2400.

Koffka, K. (1935). *Principles of gestalt psychology*. New York: Harcourt Brace.

Krueger, L. E. (1984). Perceived numerosity: a comparison of magnitude production, magnitude estimation, and discrimination judgments. *Perception & psychophysics*, *35*(6), 536–42.

Kubovy, M., & Podgorny, P. (1981). Does pattern matching require the normalization of size and orientation? *Perception & psychophysics*, *30*(1), 24–28.

Landy, M. S., & Bergen, J. R. (1991). Texture segregation and orientation gradient. *Vision Research*, *31*(4), 679–691.

Li, F. F., VanRullen, R., Koch, C., & Perona, P. (2002). Rapid natural scene categorization in the near absence of attention. *Proceedings of the National Academy of Sciences*, *99*(14), 9596–9601.

Liu, K., & Jiang, Y. (2005). Visual working memory for briefly presented scenes. *Journal of vision*, *5*(7), 650–658. doi:10:1167/5.7.5

Lloyd, S. (1982). Least squares quantization in PCM. *Information Theory, IEEE Transactions on*, *28*(2), 129–137.

Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature*, *390*(6657), 279–281.

Ludwig, C. J. H., & Rhys Davies, J. (2011). Estimating the growth of internal evidence guiding perceptual decisions. *Cognitive Psychology*, *63*(2), 61–92.

Malik, J., & Perona, P. (1990). Preattentive texture discrimination with early vision mechanisms. *Journal of the Optical Society of America. A, Optics and image science*, *7*(5), 923–932.

Mandler, G., & Shebo, B. J. (1982). Subitizing : An Analysis of Its Component Processes. *Journal of Experimental Psychology : General*, *111*(1), 1–22.

McElree, B., & Carrasco, M. (1999). The temporal dynamics of visual search: evidence for parallel processing in feature and conjunction searches. *Journal of Experimental Psychology: Human Perception and Performance*, *25*(6), 1517–1539.

Miller, J. (1981). Global precedence in attention and decision. *Journal of Experimental Psychology: Human Perception & Performance*, *7*, 1161–1174.

Morgan, M., Chubb, C., & Solomon, J. A. (2008). A “dipper” function for texture discrimination based on orientation variance. *Journal of Vision*, *8*(11), 1–8.

Morgan, M. J., & Glennerster, A. (1991). Efficiency of locating centres of dot-clusters by human observers. *Vision Research*, *31*(12), 2075–2083.

Morgan, M. J., Ward, R. M., & Castet, E. (1998). Visual search for a tilted target : Tests of spatial uncertainty models. *The Quarterly journal of experimental psychology. A, Human experimental psychology*, *51A*(2), 347–370.

Myczek, K., & Simons, D. J. (2008). Better than average: Alternatives to statistical summary representations for rapid judgments of average size. *Perception & psychophysics*, *70*(5), 772–788.

Navon, D. (1977). Forest before trees: The precedence of global features in visual perception. *Cognitive Psychology*, *9*, 353–383.

Neisser, U. (1967). *Cognitive psychology*. New York: Appleton-Century-Crofts.

Oksama, L., & Hyönä, J. (2004). Is multiple object tracking carried out automatically by an early vision mechanism independent of higher‐order cognition? An individual difference approach. *Visual Cognition*, *11*(5), 631–671.

Oyama, T., Kikuchi, T., & Ichihara, S. (1981). Span of attention, backward masking, and reaction time. *Perception & Psychophysics*, *29*(2), 106–112. doi:10.3758/BF03207273

Palmer, J. (1990). Attentional limits on the perception and memory of visual information. *Journal of experimental psychology. Human perception and performance*, *16*(2), 332–50.

Palmer, J., Ames, C. T., & Lindsey, D. T. (1993). Measuring the effect of attention on simple visual search. *Journal of Experimental Psychology: Human Perception and Performance*, *19*(1), 108–130.

Palmer, S. E. (1975). The effects of contextual scenes on the identification of objects. *Memory & Cognition*, *3*(5), 519–526. doi:10.3758/BF03197524

Palmer, S., & Rock, I. (1994). Rethinking perceptual organization: The role of uniform connectedness. *Psychonomic Bulletin & Review*, *1*(1), 29–55.

Parkes, L., Lund, J., Angelucci, A., Solomon, J. A., & Morgan, M. (2001). Compulsory averaging of crowded orientation signals in human vision. *Nature Neuroscience*, *4*(7), 739–744.

Pashler, H. (1988). Familiarity and visual change detection. *Perception & psychophysics*, *44*(4), 369–378.

Pelli, D G. (1997). The VideoToolbox software for visual psychophysics: transforming numbers into movies. *Spatial vision*, *10*(4), 437–442.

Pelli, Denis G, Palomares, M., & Majaj, N. J. (2004). Crowding is unlike ordinary masking: distinguishing feature integration from detection. *Journal of Vision*, *4*(12), 1136–1169.

Pelli, Denis G, & Tillman, K. A. (2008). The uncrowded window of object recognition. *Nature Neuroscience*, *11*(10), 1129–1135.

Piazza, M., Mechelli, A., Butterworth, B., & Price, C. J. (2002). Are subitizing and counting implemented as separate or functionally overlapping processes? *NeuroImage*, *15*(2), 435–46. doi:10.1006/nimg.2001.0980

Pomerantz, J. R. (1983). Global and local precedence: Selective attention in form and motion perception. *Journal of Experimental Psychology: General*, *112*, 516–540.

Portilla, J., & Simoncelli, E. P. (2000). A Parametric Texture Model Based on Joint Statistics of Complex Wavelet Coefficients. *International Journal of Computer Vision*, *40*(1), 49–71.

Potter, M. C. (1976). Short-term conceptual memory for pictures. *Journal of experimental psychology. Human learning and memory*, *2*(5), 509–522.

Province, J. M., & Rouder, J. N. (2012). Evidence for discrete-state processing in recognition memory. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(36), 14357–14362.

Pylyshyn, Z. (1989). The role of location indexes in spatial perception: a sketch of the FINST spatial-index model. *Cognition*, *32*(1), 65–97.

Pylyshyn, Z. W., & Storm, R. W. (1988). Tracking multiple independent targets: evidence for a parallel tracking mechanism. *Spatial vision*, *3*(3), 179–197.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*(2), 59–108.

Ratcliff, R., Van Zandt, T., & McKoon, G. (1999). Connectionist and diffusion models of reaction time.pdf. *Psychological Review*, *106*(2), 261–300.

Rensink, R. A., O'Regan, J. K., & Clark, J. J. (1997). To See or not to See: The Need for Attention to Perceive Changes in Scenes. *Psychological Science*, *8*(5), 368–373.

Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *Annals of Mathematical Statistics*, *27*(3), 832–837.

Rouder, J. N., Morey, R. D., Cowan, N., Zwilling, C. E., Morey, C. C., & Pratte, M. S. (2008). An assessment of fixed-capacity models of visual working memory. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(16), 5975–5979.

Rubenstein, B. S., & Sagi, D. (1990). Spatial variability as a limiting factor in texture-discrimination tasks: implications for performance asymmetries. *Journal of the Optical Society of America. A, Optics and image science*, *7*(9), 1632–1643.

Rudemo, M. (1982). Empirical Choice of Histograms and Kernel Density Estimators. *Scandinavian Journal of Statistics*, *9*(2), 65–78. doi:10.2307/4615859

Sagi, D. (1990). Detection of an orientation singularity in gabor textures: effect of signal density and spatial-frequency. *Vision Research*, *30*(9), 1377–1388.

Sakai, K., Morishita, M., & Matsumoto, H. (2007). Set-size effects in simple visual search for contour curvature. *Perception*, *36*(3), 323–334.

Scholl, B. J. (2001). Objects and attention: the state of the art. *Cognition*, *80*, 1–46.

Scholl, B. J., & Pylyshyn, Z. W. (1999). Tracking multiple items through occlusion: clues to visual objecthood. *Cognitive psychology*, *38*(2), 259–90. doi:10.1006/cogp.1998.0698

Selst, M. Van, & Jolicoeur, P. (1994). A solution to the effect of sample size on outlier elimination. *The Quarterly Journal of Experimental Psychology Section A*, *47*(3), 631–650. doi:10.1080/14640749408401131

Simons, D. J., & Levin, D. T. (1997). Change blindness. *Trends in Cognitive Sciences*, *1*(7), 261–267.

Sternberg, S. (1966). High-Speed Scanning in Human Memory. *Science*, *153*(3736), 652–654.

Taves, E. H. (1941). Two mechanisms for the perception of visual numerousness. *Archives of Psychology*, *37*, 1–47.

Teghtsoonian, R. (1971). On the exponents in Stevens' law and the constant in Ekman's law. *Psychological Review*, *78*(1), 71–80.

Toet, A., & Levi, D. M. (1992). The two-dimensional shape of spatial interaction zones in the parafovea. *Vision Research*, *32*(7), 1349–1357.

Treisman, A. (2006). How the deployment of attention determines what we see. *Visual Cognition*, *14*(4-8), 411–443.

Treisman, A., & Gelade, G. (1980). A Feature-Integration of Attention. *Gognitive Psychology*, *136*, 97–136.

Treisman, A. M. (1960). Contextual cues in selective listening. *Quarterly Journal of Experimental Psychology*, *12*(4), 242–248. doi:10.1080/17470216008416732

Trick, L. M., & Pylyshyn, Z. W. (1994). Why are small and large numbers enumerated differently? A limited-capacity preattentive stage in vision. *Psychological Review*, *101*(1), 80–102.

Van den Berg, R., Shin, H., Chou, W.-C., George, R., & Ma, W. J. (2012). Variability in encoding precision accounts for visual short-term memory limitations. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(22), 8780–8785.

Van Oeffelen, M. P., & Vos, P. G. (1982). Configurational effects on the enumeration of dots: counting by groups. *Memory & cognition*, *10*(4), 396–404.

Vecera, S. P., & O'Reilly, R. C. (1998). Figure-ground organization and object recognition processes: an interactive account. *Journal of Experimental Psychology: Human Perception and Performance*, *24*(2), 441–462.

Victor, J. D., Chubb, C., & Conte, M. M. (2005). Interaction of luminance and higher-order statistics in texture discrimination. *Vision research*, *45*(3), 311–28. doi:10.1016/j.visres.2004.08.013

Vogel, E. K., Woodman, G. F., & Luck, S. J. (2006). The time course of consolidation in visual working memory. *Journal of Experimental Psychology: Human Perception and Performance*, *32*(6), 1436–1451.

Wasserman, L. (2006). *All of Nonparametric Statistics*.

Watson, A. B., & Pelli, D. G. (1983). QUEST: a Bayesian adaptive psychometric method. *Perception & psychophysics*, *33*(2), 113–120.

Wichmann, F. a, & Hill, N. J. (2001). The psychometric function: I. Fitting, sampling, and goodness of fit. *Perception & psychophysics*, *63*(8), 1293–313.

Wilken, P., & Ma, W. J. (2004). A detection theory account of change detection. *Journal of Vision*, *4*(12), 1120–1135.

Williams, D. W., & Sekuler, R. (1984). Coherent global motion percepts from stochastic local motions. *Vision research*, *24*(1), 55–62.

Wolfe, J. M., & Bennett, S. C. (1997). Preattentive object files: shapeless bundles of basic features. *Vision Research*, *37*(1), 25–43.

Wolfson, S. S., & Landy, M. S. (1998). Examining edge- and region-based texture analysis mechanisms. *Vision Research*, *38*(3), 439–446.

Woodman, G. F., Vecera, S. P., & Luck, S. J. (2003). Perceptual organization influences visual working memory. *Psychonomic Bulletin & Review*, *10*(1), 80–87.

Woodworth, R. S., & Schlosberg, H. (1954). *Experimental psychology*. New York: Holt, Rinehart and Winston.

Xu, Y., & Chun, M. M. (2007). Visual grouping in human parietal cortex. *Proceedings of the National Academy of Sciences*, *104*(47), 18766–18771.

Zhang, W., & Luck, S. J. (2008). Discrete fixed-resolution representations in visual working memory. *Nature*, *453*(7192), 233–235.

Zhang, W., & Luck, S. J. (2009). Sudden death and gradual decay in visual working memory. *Psychological Science*, *20*(4), 423–428.

Zhang, W., & Luck, S. J. (2011). The Number and Quality of Representations in Working Memory. *Psychological Science*, *22*(11), 1434–1441. doi:10.1177/0956797611417006